

Historic, Archive Document

Do not assume content reflects current scientific knowledge, policies, or practices.

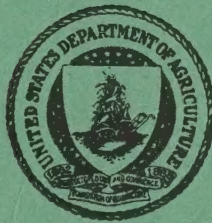
1.9
AG81ES

**STATISTICAL INFERENCE AND
THE TESTING OF HYPOTHESES**

A
series
of three
lectures and
discussions given
under the auspices of the
Graduate School of the Department
of Agriculture

by

R. A. FISHER, D. SC., F.R.S.



**U. S. Department of Agriculture
Graduate School**

December 1936

333A
LIB

UNITED STATES
DEPARTMENT OF AGRICULTURE
LIBRARY



BOOK NUMBER

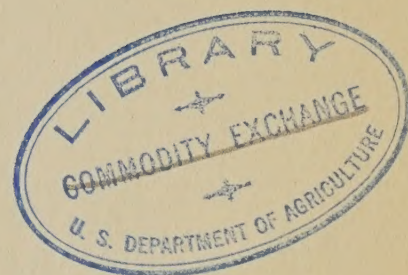
1.9

Ag81Es

483145

SPC 8-7371

Ag81Es
483145



483145

THE GRADUATE SCHOOL
OF THE
UNITED STATES DEPARTMENT
OF AGRICULTURE

presents three lectures on

STATISTICAL INFERENCE AND
THE TESTING OF HYPOTHESES

by

R. A. FISHER, D. SC., F. R. S.

Galton Professor of Eugenics
University College, London

Delivered in Washington on the
21st, 22d, and 23d of September 1936

FOREWORD

Professor Fisher's visit to Washington has proved to be a great stimulus to hundreds of investigators whose activities cover, in the aggregate, many fields of investigation. The value of Professor Fisher's lectures, and of the discussions to which he so liberally contributed, will be enhanced materially through their preservation in a written record, the preparation of which he kindly assented to in advance and left entirely to the discretion of an editorial committee appointed by the Director of the Graduate School.

The stenographers' transcript was in many places incomplete or faulty, and its reconstruction, even with the generous assistance of friends, has been carried out with difficulty and also, we fear, with flaws. The responsibility for the contents, including its imperfections, rests alone with the committee. Appreciable improvement would have been effected if Professor Fisher himself could have revised the manuscript, but for many reasons this has not been practicable.

W. Edwards Deming
B. R. Stauber
Frederick F. Stephan

Printed and sold at cost (35 cents)
for limited distribution

by

The Graduate School
of the
United States Department of Agriculture
A.F. Woods, Agr. D., LL.D., Sc. D., Director

FIRST LECTURE BY PROFESSOR FISHER
in the auditorium of the
U. S. Department of Agriculture
21st September 1936

(Introduced by Dr. A. F. Woods, Director of the Graduate School)

Ladies and Gentlemen:

Some of you may have already seen a book that I published last year entitled The Design of Experiments, and from you individually I should be glad of any indication you could give as to the desirability of my extending particular topics in this course of lectures.

For this afternoon I should like to open with a topic that is logically fundamental to the subject and also of some practical interest in view of the fact that somewhat divergent opinions have been formed by practical experimenters as to its importance. I refer to the subject of randomization. For some years I have been recommending randomization for experimental projects and though a great many people agree with me, a great many do not. Therefore it is not superfluous for me to emphasize what function it fulfills in experimentation, or rather, in that whole process including experimentation, inductive inference and statistical treatment of experimental results. These are but three aspects of one unified process by which new knowledge can be elicited.

So far as I understand it, all essentially new knowledge that can be derived by the human mind must be derived by that process. It is therefore of quite general interest, apart from technicalities concerned with the practical business of experimentation, to understand the logic by which such knowledge is obtained. Let me take a comparatively simple example to show what the experimenter means by control. We have all heard of controlled experiments and we have all heard the kind of criticism which one experimenter may levy against the experiments of another.

Let us suppose that an experimenter has injected six experimental animals such as rabbits with virulent material, which might cause death. He uses six such experimental animals and we will suppose that four of them die, leaving two. That would be a fairly convincing commonsense argument that the material used was in fact capable of killing the rabbits. It would be an uncontrolled experiment as it stands. And what the experimenter means by being an uncontrolled experiment is in particular that the experimental elements were not accompanied by a number of other animals in other respects similar but not injected with the toxic fluid. On the contrary, controls would generally be injected with serum known to be harmless.

In the uncontrolled experiment, we may feel fairly confident of a commonsense argument which would reinforce our view that the material was virulent, and it would do so because we should have a general expectation that animals treated as experimental animals ordinarily are treated, would not die over night in the proportion of 4 out of 6. We should appeal in fact to general experience or apply our own knowledge, but in doing so we should not be able to exclude the possibility that our presumption was due to some unforeseen accident such as that the water supply was poisoned that

day, or that an epidemic had suddenly broken out. Controlled experimentation aims to say: "I want to be able to say as a direct experimental fact that these animals treated in one way have been different in action from these other animals treated in another way", of which let us suppose, all six live. You notice that what the experimenter gains by that precaution is a refinement of logic rather than a reaffirmation of faith. If he had obtained this result from the uncontrolled experiment, we should, I think, continue to believe in the virulence of the fluid. A controlled experiment is aimed deliberately at demonstrating beyond a question, or at excluding a possibility that any unforeseen accident has affected, the interpretation that we place upon the data.

Now let me consider the next step of interpretation as applied to a controlled experiment represented by these data, the 2 by 2 tabulation, one of the simplest forms of statistical data that can be imagined and one on which a great deal of theoretical writing has been expended. In considering the statistical treatment of a representative body of data of this kind, we must be clearly aware of its aim. When we say that we want to ascertain whether the experiment has demonstrated the statistical significance of the difference in reaction between the test animals and the control animals, we mean that the experimental data are capable of excluding or contradicting at least at a definite level of significance--that is, at a definite degree of probability--some hypothesis respecting the reaction of the animals, and that hypothesis must be capable of contradiction by data of this kind--a hypothesis that in general we may call the null hypothesis. That hypothesis in this case is that the experimental and control animals are in fact reacting in the same way--that the two groups of animals are indistinguishable in their probability of death. We need to define a hypothesis that will lead to observably different results, such that we have calculable frequencies that we can test, and so show that it is experimentally demonstrably false. In fact, the one function of the null hypothesis as we use it in our reasoning is to supply frequencies. From this null hypothesis we attempt to infer a frequency distribution of events which we can observe to occur or not to occur. From that frequency distribution we propose to make a test of significance; namely, to recognize some small set of frequencies such as 1% or 5% of the whole. Knowing the frequency distribution, we can, of course, recognize some events as belonging to a group of classes which is rare to a certain degree in its occurrence. Those we shall choose as having contradicted the hypothesis at the chosen degree of significance. Now all that is very abstract, as logical considerations are bound to be.

Let us try to make a complete tabulation of such an easily understood experiment as we have chosen. We must distinguish the mathematics of it from the commonsense of it. Let us look at the mathematics first. Suppose p is the chance of dying under the treatment administered; then the chance of survival equals $1-p$. (Refers to blackboard). And suppose that the null hypothesis were true, that is to say, that the controls were reacting the same way as the experimental animals, or that the fluid which is possibly toxic is not toxic at all. Then the chance of dying is the same in the controls as it is in the experimental animals. If the chance of dying is p , then the probability that out of 6 experimental animals 4 die and 2 live, is $15p^4q^2$. (Referring to blackboard).

	<u>Died</u>	<u>Lived</u>	<u>Total</u>
Exper.	4	2	6
Control	0	6	6
Total	4	8	12

$$\frac{6!}{4!2!} p^4 q^2 = 15p^4 q^2$$

$$\frac{6!}{6!0!} q^6 = q^6$$

The 15 is best understood by putting 6! in the numerator for the total number of experimental animals and 4! and 2! in the denominator for the number of animals dying and living respectively. The other coefficient, unity, is similarly $\frac{6!}{6!0!}$. Then the product of these two expressions, $\frac{6!6!}{4!2!0!6!} p^4 q^8$, is the probability of what has occurred. It is assumed that p and q are the same for both experimental and control animals; that is in fact the null hypothesis.

And now you see we can take the step which is mathematically important in this argument. Notice that we do not know the value of p or q, and the hypothesis we are examining is not exactly defined in the sense that we are testing whether p or q have some predetermined values; we are merely testing whether p has the same value for the experimental as for the control animals. Yet the expression $p^4 q^8$ will have the same value for any set of results in which a total of 4 die and a total of 8 live. We may, therefore, without making any assumption or estimate of the values of p and q, compare the frequency with which such an event as we have observed occurs with that of another event which has occurred, those events being the divisions of 12 into 4 partitions in such a way as to have the same marginal values. What divisions are possible? We might have 3 dying and 3 living under the experimental animals and 1 dying and 5 living among the controls. This would be less favorable to the hypothesis; still it might be 2 and 4, 2 and 4; or it might be 1 and 5, 3 and 3; or finally it might have given a result exactly opposite from what in fact occurred, the experimental animals all surviving and 4 of the controls dying. As you can easily assure yourselves, these are the only possible results of the experiment which would give the same total number dying, living, experimental and control; that is, the same 4 marginal frequencies. The relative frequency of these 5 possible events, one of which has occurred, will be provided independently of this unknown quantity, p, representing the probability of any particular animal dying under treatment.

Theoretically, the easiest approach is to realize that if we were to recognize all these 5 results independently, all of this part of the expression, $p^4 q^8$, would remain the same, but this part, the coefficient, would vary through replacing the 4 frequencies actually observed by the set of 4 frequencies which might have been observed had the experiment turned out differently. (Illustrating on blackboard).

		Relative frequency
4 2	$\frac{6! 6!}{4! 2! 0! 6!}$	1
0 6		
3 3	$\frac{6! 6!}{3! 3! 1! 5!}$	8
1 5		
2 4	$\frac{6! 6!}{2! 4! 2! 4!}$	15
2 4		
1 5	$\frac{6! 6!}{1! 5! 3! 3!}$	8
3 3		
0 6	$\frac{6! 6!}{0! 6! 4! 2!}$	1
4 2		
		<hr/> 33

And so one can do the arithmetic fairly easily. Adding these 5 possible events together or rather their proportionate frequencies, one gets 33. So we should remember that whatever happens to the water, or when an unsuspected epidemic comes that we do not know the cause of, or when anything unknown happens that we do not know that cause of, if any such case whatever has happened of this kind, yet this observed result has only one chance in 33 of occurring among events having the same marginal frequencies.

There is a shortcut method to that ratio 1:33, which is worth asking about. I guess I can use the original table. (Demonstration on blackboard). The shortcut method is to write the factorials of the marginal values, 8, 6, 6, 4, and divide them by the factorials of tabular entries, 6, 4, 2, 0, and of the grand total, 12. It is easy to see that some of these cut out at once:

$$\frac{8! 6! 6! 4!}{6! 4! 2! 0! 12!} = \frac{1}{33}$$

Using that same shortcut, one can easily calculate the result of supposing only 3 of the experimental animals have died and 3 have survived, there being still 6 survivors from the controls. Supposing the null hypothesis to be true, we should find the probability of getting so extreme a result as we have got, to be 8 in 33. And by the convention usually adopted by experimenters, we should not be willing to assert on the basis of such a probability that an experiment of significance had been obtained. In saying that, let us be quite conscious that we are deliberately aiming at a test entirely independent--that we are not aiming at a test that rests on common knowledge; we are aiming at a rigorous self contained test which shall be in itself a demonstration of what we are hoping to show. When an experimental result is not significant and rightfully rejected as such, it is wrong to assume that it is therefore negligible as valuable experience or as a guide to our commonsense in forming opinions. It may nonetheless provide evidence which is as good as that upon which we form practical decisions such as buying stock or marrying a wife. But it isn't satisfactory scientific evidence because the opinions upon which we rested our personal opinion are based upon personal experience which we can not expect all other workers to have shared.

Now the aim of what I have been saying is to illustrate in outline and give a simple case--the line of argument essentially familiar in all experimental work by which we exclude certain hypotheses as untenable; and I particularly want to call your attention to one feature in our practical research which is recognizably essential if the argument I have developed is to be regarded as valid. It is that the assignment of individuals by these classes, experimental and control, must be strictly and honestly at random. It is very easy to see that without that precaution the result of the experiment and calculations based upon it and other care which has been expended in experimental set-up might be entirely vitiated. If the hypothesis were really true, those 4 animals that died might be recognizably sickly and if assigned to one group rather than to another on the basis of external appearance or even inadvertently on the basis of how easy they are to catch, you might very easily demolish the whole experiment. One who takes the first 6 is not making a randomization. If he chooses 6 out of 12 at random, then he knows what he is doing and knows he is not influenced by judgment or by behavior of the animals themselves. That is one example where I think you can recognize rather easily exactly what the function of randomization is. It is a supplement to guarantee the validity of the test of significance that we apply. If in fact the null hypothesis were true and if these 6 have been chosen out of the 12 entirely at random, then it falls of necessity

that this event, $\begin{smallmatrix} 4 & 2 \\ 0 & 6 \end{smallmatrix}$, will occur equally frequently with that event, $\begin{smallmatrix} 0 & 6 \\ 4 & 2 \end{smallmatrix}$, and that $\begin{smallmatrix} 3 & 3 \\ 1 & 5 \end{smallmatrix}$, will occur 8 times as frequently; that this event, $\begin{smallmatrix} 2 & 4 \\ 2 & 4 \end{smallmatrix}$, will

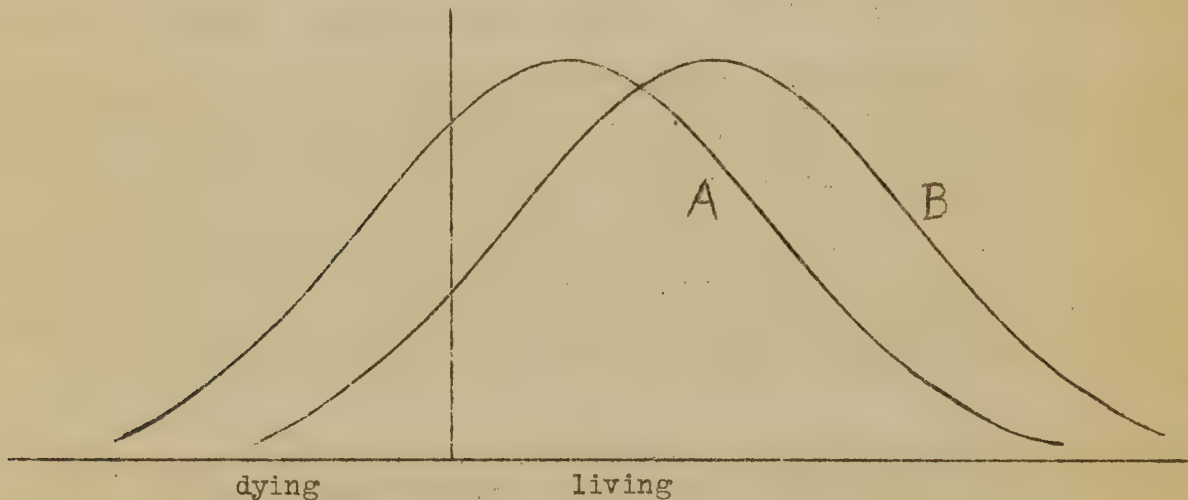
occur 15 times as frequently. Notice that those quite exact statements of relative frequency are not vitiated, however proved might be their dependence on definite causes affecting the survival of our different animals. All that we need to know to make those statements with certainty is that they have been chosen at random and that the null hypothesis was true. That last is not an assumption. It is an essential part of our argument because unless we admit that the hypothesis is true, we can not arrange the contradiction or falsity; and the whole function of the experiment is to be able to prove by statements of scientific nature whether or not the null hypothesis is disproved.

I might ask you to notice in connection with the last example just one further point, namely, that the test of significance is capable merely of disproving a particular hypothesis. It does not discuss and it need not postulate to what cause any difference in behavior between the two groups of animals may be ascribed to. It is important to a test of significance that it be free from the necessity of introducing any elaborate type of background or alternatives which might be true.

Those familiar with statistical literature will know that a great deal of the discussion of 2 by 2 tables has turned on differences in the view of it by different statisticians. For example, a well known method of interpretation, which gave rise to the tetrachoric correlation coefficient, arose from the particular view that the frequencies in the four compartments result from a normal distribution surface cut by two lines at right angles. This might happen if the classifications in the 2 by 2 table are the result of continuous variates. Thus the health of rabbits, if a scientific measure of this could be found, might be a continuous variate normally distributed, and the classification into living and dying

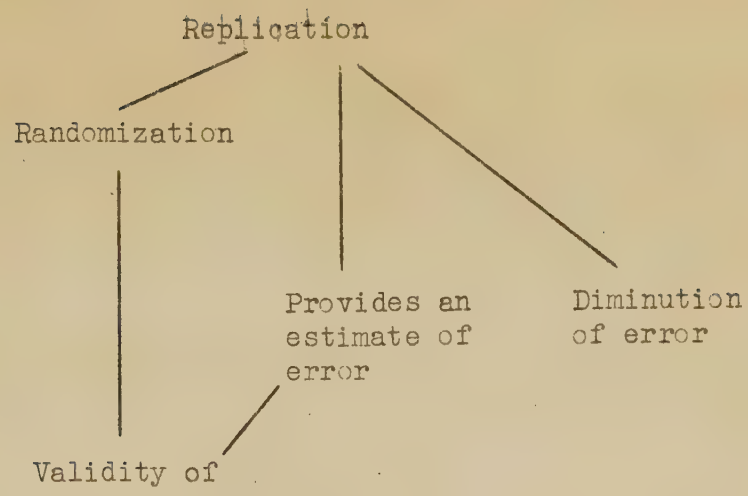
would amount to cutting this distribution at the point at which health is just great enough to sustain life. In order to have a bivariate distribution, it would be necessary to assume that both classifications are the result of continuous variates, and if both of these variates are normally distributed for all values of the other variate, we have a bivariate normal distribution. In such a case the question to be answered by the 2 by 2 table is whether these variates are correlated, and the appropriate measure for doing this is the tetrachoric correlation coefficient. The hypothesis of a bivariate normal distribution, however, is a very special one, and there are other hypotheses that are at least as reasonable in many cases.

We may, for example, consider one of the classifications as resulting from a continuous variate and the other as having one of two possible values. Thus, if the four frequencies in the 2 by 2 table are the numbers dying and surviving an epidemic among those inoculated and not inoculated, we may consider the classification into those dying and those surviving as the result of a continuous variable, as before, with inoculation or the lack of it affecting the distribution of this variable. If the distribution of the "health" variable is normal and if inoculation changes only the mean of this distribution, then the question to be answered by the 2 by 2 table is whether the means of the two distributions are the same. The situation may be represented diagrammatically as follows:



The curve A is the distribution of those not inoculated, and the curve B is the distribution of those that have been inoculated. Now the two measures that should be used under these two different hypotheses are entirely different. What I am saying is that we can strip all such scene painting from the hypothetical background if we test merely whether the probability of death among those inoculated and those not inoculated is the same.

Now let me put in rather a wider setting the function which randomization plays in providing a valid test of significance. We could put first a whole sketch which I sometimes use in relation to field experiments in agriculture. We have the primary principle of replication which holds two distinct and different purposes.



It fulfills the purpose of diminishing the error, a purpose which is being widely appreciated. As a rule, an experimenter asks for more replication. He wants to be able to try out an experiment more extensively than was previously provided for and will frequently state that his reason is to diminish the error to get a more accurate result. This purpose has been very widely appreciated, but it has another function, namely, to provide an estimate of the error.

In the very simple case of four-fold tabulation that I began with, we saw that we could judge the significance of the result, that is, of the apparent difference of reaction of two different groups of animals. We could judge that effect by obtaining a frequency curve of some observable quantity such as the number of experimental animals that died. In a quantitative way where we are concerned with quantitative measurements, we should also want to make a test of significance. In order to make this with validity, we should need to make an estimate of the error. It is on those that we must base our estimate of the precision of the comparisons we wish to make with others treated differently. Now the function of randomization, just as it was in the other case, is to guarantee the validity of that estimate. It is to make sure that areas of land treated alike and areas of land treated differently shall be strictly comparable. Or in other words, that any two areas of land which may be treated alike, shall all have the same chance of being treated alike.

Let me contrast just two methods of experimentation from that point of view. We might suppose that we have five treatments or varieties to compare in an agricultural experiment. An experimenter who would not accept the principle of randomization might apply his five treatments or varieties in a systematic order and have a number of blocks of land, as many as there are replications in the experiment.

For example, he might arrange two replications of his five treatments, a,b,c,d, and e, thus:

a		
b		
c		
d		
e		
a		
b		
c		
d		
e		

Block 1

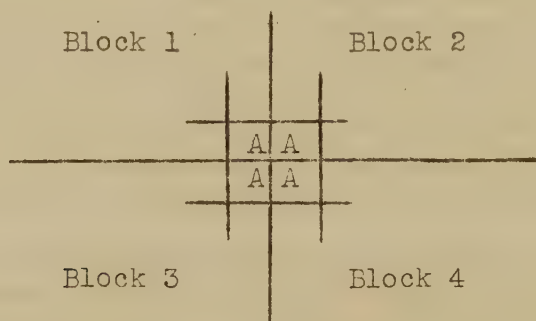
Block 2

In such an experiment certain plots will always be treated differently. Certain plots will always be treated alike. There is no question of probability as to being treated alike and being treated differently. It may be true that Nature has already randomized strips of soil but if so, it is not our doing. It is a gift from without. It may be true. If so, the experiment will be valuable.

But if we would want to rely on what we certainly do not know ourselves, we should proceed to apply these 5 treatments at random. For 5 replications, we might have a result like this:

A	E	C	B	D
B	D	A	E	C
D	C	B	A	E
C	A	E	D	B
E	B	D	C	A

(Demonstrates on blackboard). If you actually carry out randomizations of this kind, I think the first thing that strikes one who is a practical, conscientious experimenter, is the fact how badly the plots will be arranged. I mean how often such an accident as several replications of the same treatment coming together will occur. It will even happen sometimes that 4 blocks will receive the same treatment at the corners of four blocks, thus:



Realization of this possibility may come to one as a psychological shock. The strength of that psychological shock which one gets on practicing randomization is, I think, the measure of the strength with which we would, if we allowed ourselves to assign plots, by our judgment bias our experiments. We should in fact distribute each treatment more evenly than the errors, the effects of which we should like to estimate.

If we arrange them at random then these two plots (pointing to a pair of diagonally adjacent plots), which under systematic arrangement could not be treated alike, have 4 chances out of 5 of being treated differently. (Of course there is equal chance of being treated differently and the same is true of any two plots not in the same block). Consequently, the comparisons upon which an estimate of error is based, representing contrast between reactions of blocks treated alike, have always the same

and correct chance of contributing to our estimate of error as they have of contributing to real errors of our experimenting. The whole function of our experiment is to give an idea of these errors. The necessity of replication is to provide a means of estimating the experimental error.

What happens when we let ourselves go or rather when we let ourselves jeer at the consequences of randomization and say "We will have a better experiment"? That, of course, has often been done. For example, consider the following design.

Knut Vik

A	B	C	D	E
D	E	A	B	C
B	C	D	E	A
E	A	B	C	D
C	D	E	A	B

A treatment comes up once in each row and once in each column of the square. This actual arrangement was designed by Knut Vik and was a beautiful achievement in the art of distributing plots so that no two similar plots should come very close together. Each treatment is nicely distributed over the whole area. You see that in that arrangement of a square every treatment is spotted nicely over the whole area so that it is very unlikely that any of the larger variations shall affect all the plots of one treatment in a similar direction.

Now the consequences of that success are unfortunate. Let us look at it theoretically, analyzing the results of an experiment with a 5 by 5 square - one in which every treatment occurs once in each row and once in each column - as in a random arrangement square. We have a total of 24 degrees of freedom. We should eliminate 4 degrees of freedom for rows because each row bears all 5 treatments. Therefore, any general difference in fertility between one row and another will not affect the experimental comparisons. We eliminate 4 degrees for columns for exactly the same reason. Any difference between fertility between one column and another can not affect our experimental comparisons.

ANALYSIS OF VARIANCE

	Degrees of freedom	Sum of squares
Columns	4	-
Rows	4	-
Treatments	4	-*
Error	12	-**
	24	-

We then separate 4 degrees for treatments leaving 12 degrees for error to making independent comparisons at the yields of the 25 plots. Now let us see what happens if we improve upon the arrangement and are successful, as I think this systematic arrangement might be expected to be successful, in making the plots which are treated alike well distributed over the area-- that is to say less like each other in fertility than a random selection (subject to column and row restriction) would be. Its effects can best be traced by supposing the null hypothesis to be true, because one must always remember that significance is based upon calculations upon the supposition of the null hypothesis, which in this case is that these 5 treatments do not affect the yield, but are all alike. In that case, supposing it was a uniformity trial, the only thing which makes any difference in the analysis of variance must consist in the particular yields that are added together to constitute the treatment effects. I mean this; that the yields of every individual plot of a block are not altered whether we arrange the letters as in the randomized square or arrange them as in the Knut Vik square (referring to blackboard), and consequently, the total sum of squares is invariable. The same is true of the total of each row and the total of each column.

There is only one difference which can be introduced by the two different analyses, namely, that one analysis may have more of the sum of squares in the treatments and less in the error while the other analysis has less of the sum of squares in treatments and more in the error. However, in the null hypothesis, this portion which we label "treatments" represents our real experimental errors while the difference represents our estimate of those errors. If we are successful as I think Knut Vik is in making these treatment areas more representative than random selection would have been, then it is clear that this analysis has less of the sum of squares in this portion of treatments (pointing to above table at the place we have marked *) which represents the actual errors of the experiment, and consequently more in the 12 degrees (pointing to place we have marked **) which represent the estimate of error from the experiment. I should like you to see that as a matter of logical necessity.

Now I think one can judge clearly between the perfectly natural human craving to diminish the real errors of the experiment, and the logical and perhaps mechanical reliance on randomization. The square designed by Knut Vik had the effect anticipated and aimed at. You see the consequence. The real errors of the experiment are diminished. The estimate of these errors is enhanced. Not more reliance, but less reliance, is placed on the result. A result which might have been recognized as significant had the random arrangement been adopted might now be rejected as insignificant because the estimate of error has been enhanced. When you consider that a great deal of time, money and trouble has been taken, perhaps over a series of years, in order to give some effect a chance of showing itself to be significant, it then becomes a serious matter if we increase the risk of rejecting as not significant any information that might be of guidance to us. There is no safety in merely enhancing our estimate even if we have the theoretical or Platonic satisfaction that our real errors have been diminished. I think that is broadly the case for randomization, in regard to the rather subtle effects of failure to randomize.

SECOND LECTURE BY PROFESSOR FISHER

In the auditorium of the U. S. Department of Agriculture,
22d September 1936

(Introduced by Mr. Frederick F. Stephen, Secretary of the
American Statistical Association and Editor of the Journal)

Ladies and Gentlemen:

I mentioned yesterday that I had to make a selection of some topics on the general subject of the design of experiments, and this afternoon I think I ought to commence by listing and describing in brief terms the various devices that have been found successful in increasing the precision of experiments.

In my previous lecture, I devoted the whole time to a single point of, I think, fundamental importance for the understanding of the logical background of these experiments. But the function of randomization is not to increase precision. I believe it is as compatible with as high precision as can be obtained, but its function is to make one's estimate of the precision accurate and reliable. To make the errors themselves, together with the errors in the estimates of them, as small as possible, is the topic which I want to consider in the very briefest outline this afternoon; and for that purpose some five or six devices have been successful.

I will list some headings which may serve as mnemonic for those different devices.

The first trick or dodge used for increasing precision is that of pairing or grouping; next, a combinatorial trick that is best illustrated by the Latin square, or probably replications of the familiar type or principle; third, I put down the factorial design; fourth confounding, including partial confounding; and fifth, the use of concomitant measurements, used of course to obtain corrections or adjustments based on regression, along with the analysis of covariance.

These may be listed as follows:

- (a) Pairing or grouping;
- (b) Latin square;
- (c) Factorial design;
- (d) Confounding, including partial confounding;
- (e) The use of concomitant measurements; and
analysis of covariance.

To each of these topics one might be able to devote a lecture as long as I shall be able to give this afternoon; but I want you, in the first place, to consider that whole group of methods as having a single aim, the aim of increasing precision -- diminishing the magnitude of errors -- an aim, therefore, complementary to that which I was speaking of in the last lecture, which was to guarantee that whether our errors were great or small, we should have a valid and unbiased estimate of them.

Now, the first topic in that grouping is a very familiar one and can be, I think, briefly exemplified. It consists of choosing for comparison homogeneous groups, or groups as homogeneous as possible.

The experimenter is often faced with an apparent dilemma: he wishes to increase the number of his observations to increase his accuracy, because he feels that he needs to do so in order to obtain significance. He usually has only limited quantities of a highly homogeneous material. Take typical examples of that in, let us say, animal physiology. He may find it desirable to use animals closely related and therefore genetically uniform; but the supply of that material may be limited. Again, he may often find it desirable to use the same breed as being genetically more alike, but of course the number of animals in any one litter will be quite limited.

In human experimentation, one might say that the ideal material consisted in identical twins, and though large numbers of identical twins could be mobilized, there would be only two in each twinship. Then, in agriculture, it is a very old generalization that blocks or plots of land close together are more similar, and so, by choosing relatively small and compact areas of land, one can choose a group of comparatively homogeneous material. Any one wishing to compare the reactions of hospital patients to a variety of medical treatments would be wise, of course, to choose for comparison groups of patients, for example, of the same sex, of approximately the same age, and, if there are great racial differences in the hospital, of the same race. The limited amount of material belonging to a homogeneous group, of course, becomes more limited the more strictly we insist on its being homogeneous; if we are content with a low level of homogeneity, we can make it larger, or if we want a higher level, we can make it smaller.

The dodge of pairing or grouping is an attempt to make the greatest possible use of all the homogeneity that we can get by dividing our material into a number of different groups. The simplest case is that in experimental agriculture, at least it is the simplest to demonstrate on a blackboard; and one can easily imagine the homogeneous blocks of land--although in exactly the same way, the same principle applies to selections of relatively homogeneous hospital patients or homogeneous animals. Within each of these blocks of land all the different experimental treatments are applied, and so any differences in fertility between the different blocks are eliminated from the experiment. Consequently, we need only to make our blocks large enough to contain all the different treatments that we wish to compare. If we should be wasteful of homogeneity we should obtain, in fact, less homogeneity, provided we scattered all our treatments at random over the whole row of four blocks instead of insisting that each treatment should occur within one plot within each block. From that point of view, then, the application can be increased indefinitely without varying the degree of homogeneity of the material, as bringing in more blocks does, in fact, bring in land of more different degrees of fertility than when fewer blocks are used. That makes no difference as to the precision of the experiment, because once one has introduced the principle of grouping, the errors that we made would depend only on the heterogeneity within the selective groups; variation between the selective groups is eliminated.

I need not spend much time on illustrating the principle of the Latin square, which I think is familiar to most workers, and consists mainly in arranging for the elimination of strips of fertility. It is forced upon the attention of agricultural workers in particular, because of the peculiar nature of the fertility heterogeneity of agricultural land, which very often shows strips or stripes of fertility running lengthwise as a rule; broadly speaking, that is, stratified in one of the directions in which the land has been traditionally worked. What the causes of these stripes of fertility are is impossible to determine as a rule, but one knows that the land may have been manured unequally, that strips of different crops may have been grown at different times, or that it may have been laid up in lines and ridges and furrows for drainage. For a variety of causes, any land that has been long cultivated is likely to be affected in this sort of way. But it is usually impossible to determine by mere inspection or by knowledge of the history of the land whether the stripedness is likely to be more marked in one direction, say eastways and westways, rather than in the other direction, north or south. Consequently, it was early impressed upon the notice of agricultural experimenters that it would be advantageous, if possible, to eliminate the effect of variation in fertilities in both directions, and the natural solution of that problem was to start in the Latin square, in which the effect of fertility variations between rows and also between columns has been eliminated from the experimental area by making sure that each variety should occur once in every row and once in every column. One variety such as A might be placed in some such manner as this (filling in the squares of the chart below), A occurring once in every row and once in every column; and if the square is filled in with the other five letters so that the same is true of each of them, then we shall have the solution of the problem of the Latin square and such solutions clearly enable one to eliminate differences in fertility between the rows and between the columns.

A					
		A			
			A		
	A				
					A
				A	

Such solutions always exist for squares of all sizes. The most useful sizes of squares are those from 4 to 8, though the Latin square has been quite successfully applied with as many as 8, 10, or 12 to the side. It is not so conspicuously more efficient in those larger sizes; and, of course, beyond 12 one is getting into a number of replications larger than most experimenters want to use. But you can always permute the columns in the same number of ways, and finally, you can permute the letters among themselves in the same number of ways, and you might think that you will get 720^3 different squares. The actual number of different squares in any one of these permutation sets is smaller than 720^3 , but it is very large in any case.

But I shall not spend a longer time with the Latin square. It is of factorial design that I intend to speak most of the time this afternoon, so I will just pass it over now and mention that one of the features in the factorial design which leads to trouble and difficulty is that it encourages us to use a very large number of different treatments in the same experiment.

Now from what I was saying about pairing and grouping, you will see that to use a very large number of different treatments or varieties in the same experiment is to ask for very large blocks, and to have very large blocks is to have blocks that are less homogeneous than if they were smaller; consequently there is a real difficulty in precision when we attempt to take advantage of the factorial design because it increases the number of different treatments that we wish to compare; and in covering that field we might notice that it is important to handle very large numbers of different varieties.

Now the principle of confounding was introduced with a view to overcoming the difficulties encountered in the use of a very large number of different treatments and it has incidentally shown itself powerful in overcoming similar difficulties faced by those who wish to use a large number of parallel varieties, but the cases are not exactly similar and it is worth giving a little attention to the differences. Confounding then is a means--it is a device-- of choosing blocks smaller than the Latin square or factorial replications. I can illustrate that in one very simple case, which is fortunate, because nearly all the good examples of confounding are difficult. They are worth working out in detail by any one wanting to master the subject.*

Finally the last class, concomitant measurements, is familiar. There are a few who are interested in the statistical methods of these problems as introducing that very wide class of methods which we include under the title of regression. The typical instance of this kind occurs where we are measuring the success or failure of any particular treatment, such as the effect of feeds upon the weight of an experimental animal. In such a feeding experiment the final weight of a particular

* Cf. R. A. Fisher, THE DESIGN OF EXPERIMENTS (Oliver & Boyd, 1935) Art. 43.

animal is naturally connected with some other measurement taken in the course of the experiment, such as the initial weight of the same animal. Quite crudely and inevitably such concomitant measurements have always been used by experimenters. For example, if I put WF for final weight, and WI for initial weight, we might bring in the initial weight merely in such a simple formula as WF minus WI, i.e. the gain in live weight during the course of the experiment.

But equally, the experimenter might, and some would prefer to, take the ratio of the final weight to the initial weight, or possibly not that ratio, but that ratio minus 1 — it makes no difference, or what still makes no difference, the percentage increase of live weight, $100 \left(\frac{WF}{WI} - 1 \right)$. Let me say at once that there is nothing wrong or invalid in an experimenter using a concomitant measurement in such a way. If greater precision is his aim, it is in general possible to find a regression of final weight on initial weight and a regression coefficient b such that the sampling error of an expression like $(WF) - b(WI)$ shall be a minimum. There is no reason to suppose that such a compound is any less sensitive to real differences in the quality of feeding stuffs used than such a compound as $WF - WI$ or WF/WI , and in consequence we are likely to increase the sensitiveness of the experiment and are certain to increase its precision if we employ a corrective factor of this kind for the different initial weights of animals assigned to different treatments. That, quite simply, is the principle of the use of concomitant measurements, the regression technique and the analysis of covariance. One may know at once of many different concomitant measurements that may be introduced. One treating the effects on the growth of children, let us say, of raw or pasteurized milk, might presumably usefully know the age, weight, even the height, of the children concerned, because the reaction of a child to an additional nutrient might quite conceivably be affected by all those factors. You can only know by actual experiment whether the use of the factors was necessary.

And so, we should probably produce a formula of a type such as this:

$$(WF) = b_1 (WI) + b_2 (\text{age}) + b_3 (\text{height})$$

the coefficients being either positive or negative according to the actual nature of the effect upon the reacting. The use of a multiple regression formula of that kind, allowing for all three factors, might show that once we had already accounted for weight and age, we should find that the third, namely height, was negligible and unimportant, but we could not reach such a conclusion without trying.

And now, I want to turn to this question of factorial design as one perhaps most worth considering in a single lecture, with only one of these devices to be considered in any detail.

The principle of pairing and grouping and the Latin square are both concerned with the choice of experimental elements to be assigned to the treatments; and what is peculiar about factorial design and what probably delayed its recognition as a deliberate principle or policy in experimentation, is that it achieves its ends by

the choice of the treatments to be used in conjunction with each other rather than by a choice of what experimental elements those treatments shall be assigned to.

A very simple case of this kind is discussed at some length in my book,* in which four different ingredients, A, B, C, and D, allow you to make up in parallel 16 different mixtures having respectively more and less of each of the four ingredients. It facilitates discussion a great deal in this topic to get an intelligible notation - a notation that shall first, sort out the complexities and second, bring out an important contrast between the treatments and the comparisons of effects of those treatments.

If we represent some one of our sixteen different mixtures by the symbol (1), we may use the letters a, b, c, and d in designating mixtures that differ from (1) with respect to the ingredients or factors, A, B, C and D. We may combine these symbols according to the following scheme.

(1) Control	(a)	(ab)	(abc)	(abcd)
	(b)	(ac)	(abd)	
	(c)	(ad)	(acd)	
	(d)	(bc)	(bcd)	
		(bd)		
		(cd)		

(6 replications)

In the first column (1) represents the mixture that we choose to start with and which we may call the control; in the second the various ingredients are added singly, in the third column are the various combinations of ingredients two at a time, in the fourth are the combinations taken three at a time, and in the fifth column are all the ingredients combined making 16 combinations in all. It is supposed these combinations are repeated in 6 replications so that there are 96 responses in all. Within any replication the selection of treatments of course would be random.

It will be convenient to start with any one of the mixtures, which may be called the control, but it is quite immaterial. Then you may pick out another mixture (a) that differs from the control with respect to ingredient A, another (b) which differs with respect to ingredient B, another (c) with respect to ingredient C, and a fourth (d) with respect to ingredient D; but each differing from the control in respect to only one ingredient. When I say differ in respect to

*R. A. Fisher, THE DESIGN OF EXPERIMENTS (Oliver & Boyd, 1935)
Art. 38.

any one ingredient, I mean if one has a larger dose of A, then another will have a smaller dose. Again, you may certainly find mixtures which will differ in respect to two of these ingredients, as, for example, (ab), (ac), (ad), (bc), (bd), (cd), -- there are only six of these.

To combine two or more of all the different factors in all the combinations that they naturally give rise to is the whole principle of a factorial experiment, and I want to give a few minutes to say that that procedure is attended by very considerable consequences. It is more necessary perhaps to do so because it is quite contrary to a great deal of what one may call academic or logical teaching, because in a great many expositions of the scientific method from an abstract standpoint, you will find considerable insistence on the advantage, or supposed necessity, of holding constant all factors but one, while that one is being investigated; that is, of asking Nature only one question at a time. A number of similar aphorisms will be entirely familiar to anyone who has studied these expositions of the scientific method. The factorial design is in flagrant opposition to those principles; instead of asking Nature one question at a time, it insists on the filling out of quite a substantial questionnaire.

Well now, this notation, you see, is quite effective in designating or distinguishing 16 different compounds that differ in respect to four factors, and those I have represented by small letters in parenthesis. You can use these same symbols to represent not only the 16 different mixtures that can be made up, but to represent the effects of applying those mixtures. Supposing that these mixtures are concerned with a lubricant; you might measure quantitatively the friction after any of these mixtures has been applied; or they might be medical prescriptions, and we would want to observe the patient-- perhaps measure how long it took for his temperature to fall below 100°, or take some other measure of its efficacy. We will suppose then that each of these 16 mixtures has been applied to say six cases, and that we have a record of the 96 responses to these different mixtures. I say that not only the mixtures but the magnitude of the responses might be designated by those symbols. Now if we do perform the experiment in that way, we should naturally want to know whether the greater or the smaller amount of, say, ingredient A had been advantageous, and we should see, if we looked at these 16 different formulas, that corresponding to every one in which A was lacking, there would be one, in other respects quite similar, in which A was present. Consequently, if we were to compare the performance or yield of (abcd), for example, to the performance or yield of (bcd), we should be making a comparison which will be invaluable, and in which differences were due solely to the inclusion of A; and that comparison could be reinforced by several similar comparisons in each of which the contrast is due either to experimental error or to the effects of ingredient A.

As a measure then, of the success of the ingredient A in producing whatever effect is desirable, or by which the success of the mixture is to be judged, we would take the pairs of differences as follows.

$$\begin{aligned}
 & - (a) + (1) \\
 & - (ab) + (b) \\
 & - (ac) + (c) \\
 & - (ad) + (d) \\
 & - (abc) + (bc) \\
 & - (abd) + (bd) \\
 & - (acd) + (cd) \\
 & - (abcd) + (bcd)
 \end{aligned}$$

But we get a lot more than that because it may well be, and it is always to be suspected, that the effect of one ingredient is affected by another. It may well be, for example, that if A is tested in the absence of B the result is not altogether the same as when it is tested in the presence of B. We could only find that out -- we could only confirm our suspicion if we suspected it, or discover it if we did not suspect it -- by trying the effect of A both in the presence and in the absence of B, and with sufficient precision in both cases to enable a difference, if it exists, to show itself. Consequently, we must not only use A in greater or lesser amounts, both in the absence of B and in the presence of greater and smaller quantities of B, but we must make the whole experiment devote itself to effecting that contrast, and that is just what the factorial experiment has done.

If we take the sum of the pairs

$$\left. \begin{aligned}
 & - (ab) + (b) \\
 & - (abcd) + (bcd) \\
 & - (abc) + (bc) \\
 & - (abd) + (bd)
 \end{aligned} \right\} (1)$$

we have a measure of the effect of A in the presence of B; and in the sum of the pairs

$$\left. \begin{aligned}
 & - (a) + (1) \\
 & - (ac) + (c) \\
 & - (ad) + (d) \\
 & - (acd) + (cd)
 \end{aligned} \right\} (2)$$

we have a measure of the effect of A in the absence of B. For the 6 replications 24 pairs are included in (1), and 24 in (2). By subtracting (1) from (2) or vice versa we have a comparison of the two sets of mixtures. For each pair, the members differ only with respect to A. For the 6 replications, then, we should have 48 responses to mixtures in which ingredient A was present in the larger quantity, and 48 responses to the mixtures in which ingredient A was present in a smaller quantity. Subtracting the latter from the former, we should take one set of 48 numbers from another set of 48; we might well take the sum of these differences to measure the effect of ingredient A.

The first thing to notice in respect to the efficiency of experiments of a factorial character is that the effect of ingredient A measured in that way is measured with the full precision of 96 observations. It is measured with exactly the same precision as if the whole of those 96 tests had been devoted solely to the discovery of whether more or less of that one ingredient were desirable. It is made up of 48 different tests but each of those tests as to the precision of those including more of A against those including less of A involves only the ingredient A, but it is clear that an experiment of this type is quite capable, not only of testing one ingredient, but of testing the other three with the same precision. We should, of course, make a different subdivision - the B' against the 'not-B's' --but all four ingredients are tested simultaneously and it is that very fact that makes the factorial experiment so very efficient. One would get only one-fourth of the precision if of the 96 cases we should take 24 to test (a), 24 to test (b), 24 to test (c), and 24 to test (d). We might in that way get a certain precision-- but, by combining all four in the same experiment, we increase the precision of it fourfold.

You see then that the whole experiment is again utilized to see whether the effect of A is the same when B is present as when it is absent, and if it is not the same, to measure the difference. For this measure we have the contrast between the performance of 48 individuals treated in one set of ways with the performance of 48 individuals treated in the opposite set. There is one thing I think ought to be introduced beyond what I have said. We have spoken as though we had measured the effect of ingredient B on the response to A, but the measure that we have obtained in this simple case where A and B are tested at two levels - the measure that we have obtained - is symmetrical in respect to A and B. If both a and b are present in the symbol, or if neither is, we have a positive side: if one, a or b, but not the other, is present, we have a negative side, so that what might be thought of as the influence of B on the response to A is equally the influence of A on the response to B and, being a symmetrical relationship of that kind, it is called interaction AB, so that not only is the whole experiment valuable for testing the primary effects of the four ingredients, A, B, C, and D, but also the experiment gets six interactions between numbers of factors which are naturally designated by large letters. And further, by an extension of the same argument, if

you ask whether the interaction AB is affected by the ingredient C, it is easy to satisfy yourself that the answer to that question is the same as whether the interaction AC is affected by the ingredient B and also whether interaction BC is affected by the ingredient A. It is, in fact, symmetrically related to the three ingredients and may be called the interaction ABC or sometimes called the triple interaction. Whereas there are 6 interactions of two factors, there will be four triple interactions, ABC, ABD, ACD, and BCD.

And again, asking whether the interaction ABC is affected by D will produce a quadruple interaction ABCD of all four factors, A,B,C,D.

I don't know whether any of you feel that in introducing two notations, one with the large letters and one with the small, covering so much of the same ground, I am confusing the issue or using a redundant notation, but if that has crossed your mind, please notice this one thing: there is not a one to one correspondence between the 16 symbols involving small letters and these symbols involving capital letters which are only 15 in number--these represent the 15 independent comparisons or degrees of freedom of the 16 different treatments represented in the first group.

So, the first advantage of the factorial design, we will say, is efficiency - efficiency in the sense that not one question but many are answered by the same experiment and each one is answered with the same precision as if the whole experiment were devoted to it or, in other words, we make every one of the 96 trials contribute toward answering every one of the questions which can be answered.

A second advantage is comprehensiveness, because questions of a different sort can be answered. That is to say, questions answerable by these interactions are quite unanswerable so long as only single factors are tried one at a time. In fact, one may compare an experimenter who attempts to explore the causative relationships of his material by varying only one factor at a time, with an explorer who proceeds across the country along a single line of latitude, noting the altitude of all the points he passes over but not concerning himself with what is to the north or south of him. He might obtain an interesting section, but it would be baffling to the map maker to use it to discover what the country was really like.

There is a third advantage and it is a little less obvious than those two that I have stressed. We will call it "a broader inductive basis," and I think it is appreciated only when you consider that the experiments are intended to have an effective application. In fact, they mean nothing to the human race unless they have. Now, naturally that practical application will inevitably in some respects, perhaps important respects, be performed in conditions different from the experimental area in which the trial was made. Very often, and I think thoughtlessly, extreme standardization is advocated for experiments - a temperature control, a humidity control, and so on. This will undoubtedly add to some extent to the precision as applied to objects to be contrasted experimentally. They will add nothing to precision

when applied to parallels intended to increase precision by direct application, and they will certainly detract from the confidence with which experimental results can be offered as such standardization is applied to parallels. I mean, for example, that if a physical phenomenon which is observable has been observed frequently, carefully, and accurately, but has been observed only between 27.4 and 27.5° F, it is not so certain that it does occur at other temperatures.

Now, in the case of this factorial experiment, it may be of some commercial or scientific importance that an increase in the ingredient A has been followed by some measurable reaction. Our confidence that that will be so in the wider class of cases in which it will find effective application would increase if, in fact, an apparent response to A has occurred regularly in all eight of the different circumstances in which our factorial experiments have been tried. I think that is an argument that is a little bit difficult to put logically. Of course, the different circumstances that may affect the reaction to A may be circumstances other than those that do affect the reaction to A in the experiment, but personally, I should be influenced, and I think most rational people would be influenced, if they thought that the reaction to A was unaffected by such variations of the circumstances as had been systematically studied. I think a very rational confidence will be built up that the effect is to be looked for even in more widely different circumstances. Whether right or wrong, it is a human way of reasoning, and that is what is meant by saying that there is a broader inductive basis.

THIRD LECTURE BY PROFESSOR FISHER
in the auditorium of the U. S. Department of Agriculture
23d September 1936

(Introduced by Mr. Eric Englund, Assistant Chief of the
Bureau of Agricultural Economics)

Ladies and Gentlemen:

It has been suggested that I give some time in this last lecture to the theory of estimation as it has been developed in a series of comparatively recent advances in mathematical statistics which some of you may have already studied.

I shall not attempt to go into the somewhat advanced mathematics with which parts of that subject are encumbered, but rather shall try to put in plain terms some of the new concepts and technical terms which have been used to express them. Terms such as "mathematical likelihood" and "amount of information" have come into use. Many of you have heard of them, and some of you may be perfectly familiar with them, but others, I think, might be glad of an opportunity to hear at least an attempt to explain why these terms have come into use and exactly what they are intended to be used for. I have been speaking in previous lectures of tests of significance, and although I have chosen discontinuous variables for illustrating the use of such tests, it is possible that their most extensive use has been with continuous variables such as are met in the theory of errors.

The most characteristic example of this type, and one of the first tests of significance to be worked out, is known as Student's test of significance for the mean of a normal sample. Suppose that we have a series of controlled observations. Very often these will be a series of differences between animals, or plots of land, or some other experimental units treated in one way, and animals, or plots of land treated in some other way.

If experimental treatment has had any effect, we should expect to find the series of numbers representing differences, to differ significantly from zero; and it was the task in such a problem to show how a test of significance could be carried out with exactitude. The mathematical process is familiar to all, depending first of all on the calculation of the arithmetic mean and the standard deviation. Now it had been known for at least a hundred years before Student's work that if these observations were subject only to errors that are normally distributed about the true value, then their mean is also distributed normally with a smaller variance than the original observations. The practice during that century of statistical work was to calculate the ratio of the difference between the observed and theoretical values of the mean to the standard error of the mean as estimated from the data, and to use the tables of the normal distribution to find how frequently such a ratio would be exceeded. That process was apparently reasonable and was usually defended during that period on the ground that if the number of observations is large, then the estimated standard error will be very near to the true value,

and it is certain that the ratio of the difference to the true standard error of the mean will be normally distributed. In 1908 Student sought to find out exactly how this quantity was distributed for a small number of observations, and so to replace the assumed normal distribution by the actual distribution, which, as it turns out, looks very much like the normal, but differs from it in having a somewhat larger amount of frequency in the tails, and less in the shoulders of the curve. The main point is, however, that that distribution was calculated exactly, and consequently it was possible to make definite assertions of the kind that if the treatment were without effect, then this quantity would exceed a known amount with a known frequency.

Now mathematical advances are very seldom exploited immediately, and that is true of the advance made by Student. After a time, however, it became obvious that it was convenient to tabulate the probability in terms of the deviate, so that numbers could be read from tables, appropriate to different levels of significance so that one could look at a table to see how large a deviate corresponds to --for example-- 2.5%, 1%, or 0.5% of frequency.

In that set-up it was natural for experimenters to take a further step, having, I believe, considerable logical importance. Now that particular argument --that step in logic-- is called fiducial probability, and would be less remarkable than it is if it had occurred 150 years earlier. What makes it really remarkable is that during the 150 years or so prior to its development, mathematicians had endeavored to arrive at results respecting unknown quantities but had arrived at no agreement. The most notable attempt was made by an English clergyman, Thomas Bayes, whose manuscript was transmitted to the Royal Society in 1763, twenty years after the author's death. Bayes' method was developed under the term of "inverse probability." It has been the subject of a more or less heated dispute from that time to this. All the time there was latent in the material the possibility of deriving a valid system of probability statements respecting some of the parameters --statements that were entirely absent in the earlier work of inverse probability.

Now there is one caveat that must be attached to arguments of this kind. It is a principle that I think is recognizable, but which has certainly been missed by the more philosophical writers on probability, that if we make uncertain statements or inferences --I suppose we do constantly-- they would be valid only if they were based on the whole of the data at our disposal. A much later writer (Venn) in the 19th century illustrates that principle in a very simple way. He considers the statement that the death rate of Englishmen is higher in Madeira than in England, and that consequently one might expect to lessen his expectation of life from changing residence from England to Madeira. Noting that tubercular patients are said to live longer in Madeira than they would in England, he pointed out that the statement that a tubercular patient, being an Englishman, would have his life shortened by residence in Madeira was an erroneous one simply because it neglected to consider one particular item of the data, namely, that the Englishman in question was a tubercular patient. That is only an illustration of a much more familiar fact that if a statistician felt himself free to make a selection from the body of data at his disposal, then

he might apply the most regular or approved orthodox methods to arrive at any conclusion he might desire.

But there is a fundamental difference between the logic of inductive reasoning, in which we are concerned with uncertain statements, and the logic of deductive reasoning -- the logic of which we are familiar with in the study of geometry, in which the reasoner feels perfectly free to make use of any selection of the data, and in which the conclusion that may be deduced rigidly from any selection is valid without reference to any other inference that might be derived. It is in uncertain inference that we are under obligation to tell the whole truth. It is in deduction that we can use parts of the truth with perfect confidence.

Now the maximum likelihood estimates of the mean and standard deviation are in fact unique and remarkable estimates. They satisfy the condition that I hope to explain in a few minutes -- the condition of sufficiency. They possess the property that they alone -- they by themselves -- convey the whole of the information that the data possess respecting the population from which they are drawn. It would have been possible to carry through an argument of this type using an insufficient estimate such as the median, x_m , as an estimate of the mean, and Peters' formula $\frac{1}{n} \sqrt{\pi/2} \times S|x - \bar{x}|$ for an estimate of the standard deviation. Such a procedure, however, would be equivalent to throwing away an appreciable portion of the information provided by the data of the sample.

The method of reasoning called the fiduciary argument gives rise to fiduciary probabilities; therefore it depends in a rather hidden manner on an understanding of the principles of estimation. I want in the briefest way possible -- in not too technical a manner -- to explain on what principles we can judge between better and worse estimates of the same quantity.

In order to have a proper estimate at all, we must have some well defined quantities that require estimation -- that is to say, certain quantities that although unknown, at the same time are well defined in their nature. That situation arises when there are some unknown quantities such as a mean variation or a standard deviation of a normal population; or for another example, linkable between different factors; and of course the number of examples can be multiplied indefinitely. We have some properties that can determine events to be observed. We must determine the probabilities in general of every different set of events that can be distinguished.

And so in numerous situations in science we are concerned to infer the values of such an unknown quantity, as for example, the limits of the orbit of a comet, which accord with or agree with such observations as we can make. Now the material that we have to make those estimates out of consists of a series of observations, which will be a series of different values of x . Our methods of estimation can be defined in terms of operations, applicable to a series of observations. If we choose some function, as for example, the arithmetical mean of a series, it will differ in general from the true value in the population. But the first property of a desirable statistical estimate is that it shall differ from the true value less and less as the number of observations increases. It only means the larger the

sample that is taken, the better chance the statistic has of being right, and if it fails at any time to do this, then we shall not call it a consistent statistic. If

$$p \{ |T - \mu| > \epsilon \} \rightarrow 0 \quad \text{as } n \rightarrow \infty$$

the statistic is consistent. That is what is called the "criterion of consistency".

Your minds will naturally have gone ahead of that very simple requirement. I think you will say that we can not be satisfied merely with the requirement that a statistic tend more and more to give right values as the sample is increased, but that in the size of sample that it is derived from, it should have, in addition, smaller errors than other statistics that might be suggested.

Now it nearly always happens (and you can ignore cases where it doesn't happen) that after the sample is increased sufficiently in magnitude, the distribution of an estimate tends to a normal distribution, and what is more, that the variance of our estimate--say large T --falls off in the limit inversely as the square root of the size of the sample taken. (Demonstrates on blackboard what the formula of probable errors involves).

Estimate T	$n v(T) \rightarrow A$	$[v(T) \text{ means "variance of } T"]$
Mean	$n v(\bar{x}) \rightarrow \sigma^2$	
Median	$n v(x_m) \rightarrow \sigma^2 \pi/2$	

It is natural to prefer methods of estimation that make the chance of error, in the limit, as small as possible, and that is what is called the criterion of efficiency. Efficiency requires that at least in the limit for large samples (which is a condition necessary to the exactitude of the ideas I have been developing) the estimate we make shall have the smallest possible sampling error.

All that is very abstract, and it begins to meet practical needs only when we ask the further question, "Do we suppose there must always be some efficient statistic?" That is to say, that some must have a smaller sampling error in the limit than others? Some of them must have the minimum possible. How can we find--how can we make an estimate--so that it will have the least possible error in the limit? It has been shown that the method of estimation which satisfied that requirement is known as the method of "maximum likelihood".

As an example consider a problem of linkage in children. Suppose we made a schedule of all the possible occurrences. It is convenient to think of them as a discontinuous series. Possibly it is accurate for continuous variation as we have been speaking but let us consider discontinuous series. We observe a family of children and classify them in various ways. There are 20 different sorts of families distinguishable. According to the value of our parameter θ (the linkage), perhaps each kind of family has a definite probability p . (Demonstrates on blackboard).

Family type	Probability*	Frequency
(i)	p_1	a_1
(ii)	p_2	a_2
(iii)	p_3	a_3
⋮	⋮	⋮
⋮	⋮	⋮
(xx)	p_{20}	a_{20}

*Calculable from an assumed value of θ

We are going to try to judge the family linkage by the frequency a with which each of these different types of families is in fact observed. The probability that the observed distribution of families would occur if the families are chosen at random is

$$p_1^{a_1} p_2^{a_2} \dots p_{20}^{a_{20}}$$

We call this the "likelihood function". That is a natural and convenient definition. Then we take the logarithm of the likelihood function and get

$$a_1 \log p_1 + a_2 \log p_2 + \dots + a_n \log p_n = S(a \log p)$$

To maximize the likelihood function we set its derivative

$$S\left(\frac{a}{p} \frac{dp}{d\theta}\right) = 0$$

and solve this equation for θ .

What has been demonstrated is that if we use as our method of estimation that value of θ which causes the above derivative to vanish--i.e., which maximizes the likelihood function--then this value of θ , this estimate, will satisfy the condition of efficiency.

A second question is, "What will be the limiting variance of such an estimate?" And there we come to one of the most fundamental ideas of the whole theory of estimation. If i denotes the "quantity of information" in the limit then one can calculate

$$i = S \left\{ \frac{1}{p} \left(\frac{dp}{d\theta} \right)^2 \right\}$$

We can calculate the precision of the most precise estimate directly from the primary information on which our method of estimation is based, rather than the specifications of the probabilities with which each observable event occurs.

I can say only a word or two further, and I am afraid I have taken up too much of this lecture with rather heavy matter. Perhaps the most useful word I can add is that we have secured a good quantity or amount of informative data through the intricacy of statistical methods, and methods of combination of observations, or methods of reduction of data used by a statistician, with a given form and given body of material. But you will see that we have done something else incidentally.

We have now obtained a measure of the amount of information in the material, or the maximum precision that the statistician can possibly attain in interpreting or making estimates from that body of data. It is of some importance to have done that because it changes our outlook on the question of statistical application generally.

There was a time when a statistician could rather flatter himself and feel very much pleased if he invented a method of investigation or a statistical method in general that was more precise than someone else's method that was previously used. That system of ideas is the language of alchemy. The alchemist claims that a given person working this material can make gold out of it and the statistician who treats data in the same way is in a pre-scientific condition. What we do, when we advance as far as chemistry, is to make an assay to discover how much value the material contains, and then we are in a position to discuss the fruitfulness of the methods of extraction that may be proposed. But the essential point is that the amount of information therein is limited, and no statistical worker can increase the amount that is there. The task of the statistician is to find out how much there is and extract as much as is worth while.

But a much more important thing than that arises. Statisticians could argue, dispute, and put in a great deal of work, increasing the proportion of information that they extract from the data from 99% to 99.5%; but whatever they do they can't increase it very much more.

A study of the data from the point of view of why it doesn't contain more information than it does, or what is needed to make it contain more information, is infinitely more fruitful. Data themselves could often be increased in value about 10-fold simply by a rearrangement of the experiment. The disproportion between the fruitfulness of the experiment and what might in fact have been attained is rather striking, especially now that we know how easy it is to extract the whole of the information from any given data. Perhaps it is the intensity of my feelings for that disproportion which has made me direct my interest rather toward the design of the experiments -- the creative side by which the new information is actually produced -- rather than to pursue very much further the problem of estimation, which I still confess a great fondness for; but I don't want to suggest or pretend that any enormous amount of practical value is likely to flow from any further investigation that I can make in the theory of estimation.

A DISCUSSION ON STATISTICAL PROBLEMS

held at the Cosmos Club in Washington the 22d of September 1936

(Mr. B. R. Stauber, chairman)

MR. STAUBER: Those of us who work in the field of statistics in any of its branches are extremely fortunate in the fact that we share a community interest in a tool that is widely used in a number of other fields. We are therefore able to enjoy an opportunity to discuss our problems across the table, especially when we can share the counsel of one who has made extraordinary progress in his own field. Dr. Fisher has very kindly consented to talk over various problems that we have come across and to give us the benefit of his own experience in regard to them. Before asking questions, however, I think perhaps there is another phase of his visit in which we shall be interested. We have all had the opportunity of seeing Dr. Fisher; perhaps we are not all aware that Mrs. Fisher is now in this country with him.

MRS. FISHER: Thank you. I am very proud, of course.

MR. STAUBER: There are no doubt a number of us who have questions we should like to raise. Some have thought of questions ahead of time and have indicated their desire to raise them, but we do not want at all to restrict the discussion to any one topic. However, it is entirely likely that one question will lead to another and lead to a well-rounded discussion. Mr. Sturges, would you care to state your question?

MR. STURGES: Economists are particularly interested at this time in the correlation of the time series. Suppose we have a time series, in particular one of annual prices. If we correlate the price one year with the price the next year, on the average we get a correlation of upwards of 0.5. In other words, our annual prices are not independent of each other. There is a certain degree of relationship and it is not at all difficult to see why this should be so. Take the case of eggs, for instance. The price of eggs this year is going to be a little bit higher than otherwise because of the drought of 1934, and in 1938 and 1939 prices will be higher because of the drought of 1936. In judging significance, it is customary to consider that every year's figure that we have constitutes an independent observation. As I said, when we actually refer to a year, we find that it is not independent. Going back to egg prices again, if I had 15 years of egg prices in a correlation with three other factors, following the customary method 4 degrees of freedom would be lost, leaving 11 with which to test the adequacy of the hypothesis. Now actually, if these egg prices are independent only one year in three, we do not have 15 observations at all; we have only 5. When we subtract four degrees of freedom we have only one degree left over with which to test significance. The point I should like to raise is whether or not it is proper to use so severe a test in judging the significance of the correlation, particularly when we bear in mind that we shall wish to make inferences later on. It is of great importance that we find out beforehand whether or not a correlation is real or not.

DR. FISHER: I think Mr. Sturges has raised an extremely important point and one that gives some little embarrassment because he has said almost all that seems worthwhile to say about it. The error, the danger that he alludes to, is very widespread; also it is fairly familiar, a good many people having pointed it out. For example, the first reaction from that wave of confidence in correlation which flowed over the world about 21 years ago, began, I think, by someone pointing out that time series did sometimes give some very unreliable correlations. You might take a series such as the number of apples imported into the United Kingdom during a period, such as the last 30 years, during which the number has been increasing very largely with the gradual exploitation of the discovery that the English public is willing to buy larger amounts of fruit if it gets the opportunity; and you might correlate that series with, let us say, butter imports into the United Kingdom, which during that period have been falling very rapidly; and the correlation will be high enough to be statistically significant. I think that was one of the most useful and strikingly emphatic examples of how misleading a correlation analysis might be; and people differ, I think, only in the different names that they give to the cause of its being misleading. Mr. Yule, I think, gives a name for the cause; he calls them "nonsense correlations". That is all right if you know a nonsense correlation by instinct. Personally, I should say that the cause was essentially one of heterogeneity. Suppose we take the left and right hands of a number of individuals; take some measurements of the left hand, and corresponding measurements of the right hand, with a view to discovering how closely correlated the two hands are. Let us measure them in centimeters, but, by some clerical error, let us say, let one pair of hands be measured in millimeters. One has then a record of 130 or so, when one ought to have about 13; both hands the same--one giant pair. Then of course, anyone using these figures for correlation purposes would receive, you will recognize, a very enhanced correlation from that clerical error. The reason for that is very simple, one pair of hands does not belong in the series at all. It is not in the same picture as the others. But I say this is the same phenomenon as the time series. In the case of the clerical error involving one pair of measurements, the one degree of freedom that is heterogeneous from the others is the comparison between the giant pair of hands and the others, the comparison of the others among themselves being quite normal and homogeneous among themselves.

One approach that seems to me relevant to nonsense correlations is the removal of the linear component. Sometimes the removal of the linear component will leave the rest homogeneous. Actually, time series are habitually not absolutely homogeneous, and may be heterogeneous in more than one way.

Suppose we have a long time series, 15 years by months - 180 observations. The series might consist of a few principal recognizable types. The first type I think is one in which it is all homogeneous after the first few. That is to say, if there is a trend, and the 1st, 2d, 3d, 4th, 50th degree, or perhaps all of the fluctuations are homogeneous, then in that case, of course, the deviations from the trend line

might be non-correlative but they would show such correlations as are slightly negative between neighboring points. But it might be different from that; it might be you had a series like trade cycles repeating six wave-like movements, but shall I say descriptively, very smooth individually. In such a case you would not remove all the heterogeneity by curves of the first three, four, or five degrees; you would certainly have to have curves of degree as high, or rather higher, than the number of maximum and minima altogether; you might well find that the bulk of the variance lay between the 10th and 20th, not much before 10 and not much after 20.

I will speak as a physicist would do of genuine harmonic curves, because the amount of labor in dealing with them on this system would, I think be intolerable for most workers, and this is tough enough, but I think one can usually, and I think it is worth while to make up one's mind which type it is. Is it a type in which all the heterogeneity lies in the first few degrees? Is it a type in which there are a lot of very smooth, very minute degrees of freedom? If so, the proper remedy, I think, is as Sturges suggested. These degrees of 130 to 180 simply do not come into it--they are dummies.

MR. STAUBER: A very suggestive line of discussion. Are there any other questions that anyone wishes to raise along that line?

MR. STURGES: I did not have in mind the question of trend. What I did have in mind was just primarily the sort of magic feeling, that most of us have that a calendar year or crop year, whatever happens in 12 months, just naturally has to be independent of what happens in another 12 months. Often we find a trend line of a moderately low degree, but one nevertheless, the deviations from which will still be quite highly correlated.

DR. FISHER: It depends on how far you go. They will go negative after a bit.

MR. STURGES: For instance, take my 15 years of eggs. Suppose we take the 5th degree. Even so, the deviations from the equations will still tend to be correlated. If one year is high, the next one will perhaps be not quite so high.

DR. FISHER: That means that it is very smooth; that the actual contour of the curve is smooth, and these down here, from 10 to 15, are probably quite abnormally small.

MR. STAUBER: Are there any other points in connection with that general question?

DR. DEMING: I do not know whether I've got the point here. Supposing that instead of calling a point for every year, Mr. Sturges, you take it every month and make 180 points. Is that anything in the nature of what you had in mind, or was that extreme?

MR. STURGES: Suppose we had gone further and taken a point every day. We would have a very smooth curve for 15 years. To the eye it would be just a smooth line not necessarily with regular and recurring cycles.

DR. FISHER: You would be getting a discontinuity of the price units if you went as far as that. It could not change less than a cent for a dozen eggs.

MR. STURGES: It quite often happens in our economic series; there is no regular recurrent wave to it. Suppose we take a series beginning in 1920; in the first 10 years there is a gradual swing upwards; we come down in a big dip and then start up a bit. It would respond to an equation of high degree.

DR. L. H. BEAN: I wonder if Mr. Sturges' question does not come down to this. Suppose one of the variables is the course of egg prices; presenting a cycle during the course of the year, i.e., 12 months in each cycle, with a corresponding behavior of supply. Prices during the six months of the last half of the year will correlate almost perfectly with the prices during the first half of the year, because you have in one case a rising series of prices, and on the other a declining series of prices. The cause of the price behavior is demonstratively the variations in supply, and Mr. Sturges apparently suspects that supply is correlated with price and apparently would give a very high correlation in this particular instance. This is perhaps invalidated by the fact that the first half of the price series is highly correlated with the last half, a correlation which is basically due to a similar situation in the factor which causes the prices to vary. I wonder if economic results of this sort are invalid because of the mere assumption that where there is interdependence in successive items of the independent factor, and at the same time a relationship within the items in the dependent factor, it is traceable not to itself but to something else.

DR. FISHER: I can't say anything very clear about that. When we think in terms of cause and effect, naturally we think in terms of time sequence, so in a sense, to say there is a continuous mix in cause and effect in time is to say that a time series may well be heterogeneous in the way that I have been trying to describe. I prefer that form of statement; that is to say, that the successive degrees, or, if you will, the components of successive frequencies, are at least liable to be heterogeneous in time series because that does point to a manageable method of grouping groups of neighboring degrees. The significance, after all, is nothing but consistency. Supposing that from No. 11 to No. 20, two series had the same sign for each pair of terms, that is, 10 times running, there would be no need to calculate any more complicated test of significance. Those have demonstrated their consistency in that range of components. The fact that No. 1 is at the same side does not prove any consistency at all; it is a thing that has a half chance of being true anyway.

MR. STURGES: Suppose we get back to the game of chance. We are going to form a series of numbers by throwing a single die. Each throw I make will be one observation in my series. I throw once, but instead of making a second throw for my second observation, I will toss a coin, and I will make an independent throw only if my coin comes head. Many of our time series are of that nature and only every so often do we really get a new variation. Quite frequently the year is just the figure we had but only with some minor change.

DR. FISHER: Good. I ought to qualify what I said: it was quite empirical and without a constructive theory. If one can put forward a constructive theory, that is definitely a superior approach.

MR. STAUBER: Perhaps it would be interesting for us to pass on to a further series of problems. I believe Dr. O. A. Pope has a question or two he would like to raise.

DR. O. A. POPE: My question concerns primarily the proper way of constructing a combined analysis in dealing with certain percentage values. If we take, for instance, a number of cotton variety studies at different places in the belt; and if we determine the percent of 3-lock, 4-lock, and 5-lock bolls on each plot; then in making the combined analysis, is it appropriate to treat percentage 3-lock, 4-lock, and 5-lock bolls as independent variables? In the combined analysis there is no contribution for varieties; none for locations, and none for varieties with locations. Should such data be treated in some manner other than through percentages?

DR. FISHER: I have had a little time to look at this example, and have one or two points to make. I think I should agree that one could take at least two of those three percentages as functionally independent, but when I call them functionally independent, I must at once qualify that remark. First, as statistically correlated, they should be treated as dependent variables, that is to say as a thing you are analyzing, let us say rows, columns, varieties, places, interconnections between those things; and I should not introduce the percentage of 3-lock bolls, 4-lock bolls, or 5-lock bolls as a subdivision analogous to the difference between one variety and another. That would be treating them as independent variates rather than as a result of the experiment. I wonder if I made that sufficiently clear? We use the contradistinction independent and dependent variates for the things we know and the things we want to predict. If you look for cause and effect, they often are but ways we are seeking for some means of expressing our expectation in respect to the dependent variates in terms of our observation of the independent variates, and I should certainly treat these percentages of 3-, 4-, and 5-lock bolls as dependent and study their dependence on variety, place, interconnection between variety and place, after eliminating details like rows and columns.

There is another feature in this experiment that is worthy of comment, namely, that it was an arrangement of what is called the semi-Latin square, which for a few years after the days the Latin square became

widely recognized was put forward in quite a variety of different places all over the world as a possible means of expanding the application of the Latin square to cases in which we have more than 8 varieties. The arrangement used consists of 8 columns each of which contains all sixteen varieties, and of 8 double rows each of which contains 16 varieties, so that if you look at it from another standpoint, as an 8 x 8 Latin square with split plots, there is a distinction which I think is now clear but which was not clear when the semi-Latin square was at first widely advocated, and that is the distinction between the cases in which variety A is a component, the variety B in one double plot in the square, and where there are two varieties remaining married throughout the whole square, and cases in which variety A has 8 different components in the different double plots of the square.

If the first practice is adopted, one does have a perfectly definite and intelligible analysis, and I think that is worth putting down: first, analysis between double plots; there we have an ordinary Latin square analysis; rows, columns, treatments, varieties; error a 42 for the 63 comparisons between the 64 double plots. Next, then, double plots, or between the plots of the pair, there will be, of course, 64 comparisons, one within each double plot, and if the varieties are tied together in pairs, there will be just the 8 varieties for comparisons, varieties 8, that is to say A versus B, C versus D, E versus F, and so on for the 8 pairs of plots which are tied all together in bundles of two, and a remaining error 56, which we can call error b. So that in that form of analysis we have divided the 15 degrees of freedom among the sixteen varieties into two parts, one representing 7 comparisons, A and B together, C and D together, or any others of those 8 pairs of varieties and these 8 individual comparisons within those pairs. To the first set of comparisons, the error based on the Latin square will be applicable. To the second set of comparisons the error based on differences between the double plots will be applicable; but if we mix the thing up--if you allow a variety to have different components in the different double plots of the square-- then every varietal comparison you wish to make will be straddled between error a and error b, and you would have considerable difficulty in finding to what extent it was affected by one set of error, and to what extent by the other set; and if one assumed all error formed by pooling those two, throwing them together as 98 degrees of freedom will be definitely biased.

So you might say that the split plot Latin square is a good and useful arrangement, but the semi-Latin square in the sense in which it was originally advocated is not so good. I may say that at present. In the last year or two much more effective methods than we previously have had have been developed in comparing large numbers of varieties, and I fancy that that is a problem which concerns a good many people here.

MR. SCHUMACHER: Suppose the observation of your 8 x 8 Latin square be B-A, C-A, D-A, and so on; rather let us say that each plot contains A always and the other half is either B, C, D, E, F, G, H, or I.

DR. FISHER: Is this the case you are thinking of, one in which the 16 varieties are themselves due to the combined different factors, one

represented by the contrast C-A, and the other by some contrast such that to every one of these 8 there corresponds one of the other 8?

MR. SCHUMACHER: It is this; that one of the two halves of each plot be the control variety.

DR. FISHER: The same one all the time?

MR. SCHUMACHER: The same one all the time. Then I should like to ask about an analysis of the Latin square, the observation being the difference between A and the other half, such that your rows, columns, and treatments each contain 7 degrees of freedom, and the error contains 42. How does something like that differ from such as you have there?

DR. FISHER: If you had each of these with the same control, then this part of the analysis becomes totally unnecessary.

MR. SCHUMACHER: Might there not be two sets of analyses, one such as you have in error a, the second, such that the observations be of the differences between A and its mate, with error b?

DR. FISHER: If you had A against some control, B against the same control in an 8 x 8 Latin square, this would be a valid analysis without anything further, but I don't know whether rows and columns would be any good to you in that case because you have already eliminated the effects of rows and columns in making comparisons against controls occurring in those very rows and columns so that the amount of squares you would throw out here might be quite trifling.

MR. SCHUMACHER: Under such conditions we may have A contrasted against something else, call it the difference between control and the variety mated with control with 7 degrees of freedom, and 57 for the error?

DR. FISHER: Yes, I imagine you may. That would be 8 comparisons; that would be 8 for varieties, and 56 for error; that is if you ignore rows and columns altogether.

MR. SCHUMACHER: It would be 8 sets of differences with 7 degrees of freedom for those 8 differences?

DR. FISHER: The 7 degrees of freedom among the 8 differences B-A would be pure error. The remaining one would be for varietal difference. Other sets, C-A, D-A, etc. also contribute one for varietal difference and 7 for error, giving in all, 8 for treatment and 56 for error.

MR. STAUBER: Are there any other questions?

DR. POPE: I wonder if I might inquire what is the best way to set up field experiments in which it is desirable to compare yield among a large number of varieties, say 100 to 200 varieties. What is the most precise method of setting up the field arrangement in order to determine the differences between varieties?

DR. FISHER: There is quite a range of methods appropriate to different practices. Perhaps I can start with a very pretty arrangement which Yates has recently been developing called incomplete randomized blocks. I won't deal with the statistical analysis because it would take an unnecessary length of time.

Suppose we have v different varieties. We are not going to have v in each block because v is a greater number of varieties than we wish to put into a single block. We shall put k varieties in each block. Then if there are r replications, it is clear that we shall have vr/k blocks; and this number must be an integer, call it b . This arrangement will be most useful if we can arrange it so that every pair of varieties comes equally frequently in the same block, and so introduce that element of symmetry that will enable the comparisons to be made with the same precision. Hence $\frac{1}{2}v(v-1)$ is a factor of the total number of comparisons $\frac{1}{2}brk(k-1)$, whence $pv(v-1)=bk(k-1)$ where p is the number of pairings of any two varieties. Since $vr=kb$ (see above), we find that $p(v-1)=r(k-1)$.

We can always satisfy these equations if we take as many blocks (b) as there are ways of choosing k things out of v , which means that

$$b = v! / k!(v-k)!$$

The number of replications (r) may be determined by substituting vr/k for b in the value of b just written. This substitution gives

$$r = (v-1)! / (k-1)!(v-k)!$$

If we further ask how often any two varieties occur in the same block, the value of r just found may be substituted in the equation involving p , which results in our finding that

$$p = (v-2)! / (k-2)!(v-k)!$$

If $v = 31$ and $k = 6$, then

$$b = 31! / 6!25!, \quad r = 30! / 5!25!, \quad p = 29! / 4!25!$$

the values of which would be exorbitantly large.

Now that is certainly a possibility but is practically an inconvenient one usually owing to such very large numbers. The interest is in having fewer replications and still satisfying conditions. The arithmetical requirement is that b , k , and p have a common factor; they can all then be reduced by a such common factor, the larger the better.

Referring to the example above, the value $29! / 4!25!$ for p is, in fact, this highest common factor of those three numbers, and we have an arithmetical possibility. By this common factor p , the number of blocks of 6 items each becomes 31; r , the number of replications, 6; and p becomes 1, every variety occurring once and only once in conjunction with every other variety.

If $p = 1$, then $k-1$ is a factor of $v-1$, and $r = (v-1)/(k-1)$; also $k(k-1)$ is a factor of $v(v-1)$. The relationship $v-1 = k(k-1)$ is obvious, and $v = k^2 - k + 1$. Then $r = k$ and $b = v$. Now that is only an arithmetical possibility so far, and one has to consider the common factorial problem as to whether it is actually possible. In this case it is, and I chose that case because it represents one of the most important series of combinatorially possible solutions of this kind. Suppose we represented our 31 varieties by letters, such as a, b, and c. (There would not be enough in the alphabet--we shall need to start another alphabet to get 31.) Now suppose I make a block out of six of the 31 varieties; there are now 25 left. I am going to have 5 more cases in which a appears, and in which b, c, d, e, and f are out, so that the 5 components of a are 5 out of the remaining 25 varieties, a difference of 5 in each case. There are going to be 5 more blocks with b in them and again a, c, d, e, and f are out of it, and the components of b in those 5 blocks will have to be another subdivision of those remaining 25 in the 5 sets of 5 each. The thing will therefore be possible if we can divide up 25 objects into 5 groups of 5 each in 6 different ways, a, b, c, d, e, and f, in such a way that no two objects come twice in the same group. We can divide up our 25 objects into 5 groups of 5 in one way, and in the second way, fulfilling the condition that no two objects in the same group once can be in the same group a second time. If we can do it a third way, we can do it a fourth way, and with any prime number, we can do it in all six ways, so that this method of experimenting with a particular number in 31 different varieties in blocks of six is quite feasible.

Now, it is worth while to see what is possible and what is not. We took 5 and 31 that time, 6 blocks of 6, and 31, 31 being $6(6-1)+1$. Five blocks of 5 and 21, 4 of 4 and 13, 3 of 3 and 3, satisfy the above, $v = k^2 - k + 1$, $r = k$ and $b = v$. But our interest extends to the higher numbers, 7 and 43, 8 and 57, 9 and 73, 10 and 91. These are all known to be possible. No one knows anything about 11; that is to say, it is open to anybody to produce such a completely orthogonalized square.

A DISCUSSION ON STATISTICAL PROBLEMS
held in Room 4090 of the Department of Agriculture the 23d of September 1936

(Dr. W. Edwards Deming, chairman)

DR WAUGH: Given: The zero order correlation matrix,

$$1) \quad R = \begin{vmatrix} 1 & r_{12} & r_{13} & \dots & r_{1n} \\ r_{12} & 1 & r_{23} & \dots & r_{2n} \\ r_{13} & r_{23} & 1 & \dots & r_{3n} \\ \dots & \dots & \dots & \dots & \dots \\ r_{1n} & r_{2n} & r_{3n} & \dots & 1 \end{vmatrix}$$

based on N observations of n variables.

In terms of cofactors the regression coefficient and its standard error are

$$2) \quad \beta_{ij} = - R_{ij} / R_{ii}$$

$$\sigma_{\beta_{ij}}^2 = R_{ii} R_{jj} / N R_{ii}^2$$

where $R_{ii} R_{jj}$ is the determinant that remains after the ith row and column, and the jth row and column, have been deleted from the matrix R.

As Frisch has shown in his Confluence Analysis, both β_{ij} and $\sigma_{\beta_{ij}}$ will approach the indeterminate form 0/0 in cases of multiple collinearity. However, if the cofactor, R_{ii} , is significantly different from zero both formulas in (2) should be applicable. In all cases $0 < R_{ii} < 1$, since R is positive definite. If it is close to unity there is little correlation among the independent variables. If it is close to zero there is almost perfect correlation among the independent variables.

Question: Can you give a test of the significance of an observed correlation determinant, such as R or R_{ii} , which will indicate whether or not it is significantly different from zero?

DR. FISHER: The difficulty, I always feel, is that if these denominators do not actually vanish, that is to say, if they are not zero itself, they will always be significantly different from zero. What does "nearly

zero" mean? You can say that one thing is small in comparison with another. But if anything differs from zero, it cannot be compared with zero. Frisch's problem from the point of view of multiple space, in these coordinates, is the angle between two points on a hypersphere. The vanishing of any of these determinants is in relation to linearity. Let us take three points on a sphere, such as Washington, New York, and Boston. They lie on a great circle; yet if you chose to name three points in these three cities, they might not be exactly on a great circle. If they were on a great circle, exactly, there is an observational difference, and the question of the significance of that seems to be that it can be considered only so far as meaning can be attached to the statement whether it is significant or not. If in a population of which this is a sample, the corresponding points were exactly on a great circle, is there any probability that in the sample they should not be exactly in the great circle? Of course, from that point of view, there is no chance. If the agreed upon quantity vanishes in the population, it will vanish from every sample you take from it. In such a sense the determinants that you speak of will vanish. If a certain quantity does not vanish exactly in every sample, it does not vanish in the population.

Perhaps an analogy in linearity will illustrate that further. If we have x and y as observables and have a small sample of 3 observed points that do not lie exactly on a straight line, then we can reject very easily the hypothesis that all samples of observed points must lie exactly on a straight line; and in that sense the matrix would deviate significantly from zero.

DR. WAUGH: I wonder if there might not be some test similar to the z test. In some correlations of economic data that tend to work out this way there is a test of closeness to zero. You can set up problems in such a way that the denominator becomes pretty small, and it is quite possible that the determinant, which like the correlation coefficient itself cannot go below zero nor above unity, is different from zero only owing to errors of observation.

DR. FISHER: If you had an independent measure of observations such as the psychologists try to obtain by duplicate tests of the same individual, then you can say this does not differ from zero by more than could be explained by errors of observation of this magnitude. But I don't think Frisch has explicitly introduced any such ulterior measure. It seems quite essential to the application of his measures and it is purely qualitative.

MR. FRIEDMAN: In computing regression equations from a large number of observations, it is frequently a great saving of time to group the observations by the independent variable, compute the means of both the independent and dependent variables for the classes so obtained, and correlate the means. If the observations are drawn from a single correlated universe, or more particularly from a multivariate normal universe, the regression coefficients have the same expected values whether computed from the means or from the individual observations. This is true for the

value, it seems to follow that the correlation coefficient computed from the means will have an expected value higher than the correlation coefficient computed from the individual observations, although it will tend to be less significant. The question that suggests itself is whether it would be possible to improve the estimate of the regression coefficient obtained from the means by utilizing the knowledge of the intraclass variance of the dependent variable. I think the reason why it seems reasonable that this should be the case might be shown on the blackboard very simply.

Consider the case of two variables. Then, assuming that we have a regression equation computed from the individual observations, the total variation in the data can be divided into three sums of squares: The first, attributable to the regression; the second, to the variation of the means about the regression; and the third, to the variation of the individual observations about the means. This can be represented by an Analysis of Variance table:

Source of Variation	Sum of Squares	Mean Squares
Regression equation	A	
Means about regression	B	V
Observations about means	C	V

If the regression equation computed from the means is the same as that computed from the individual observations, then the square of the correlation coefficient as computed from the means will be given by

$$\frac{A}{A + B}$$

Similarly, the square of the correlation coefficient as computed from the individual observations will be

$$\frac{A}{A + B + C}$$

Since the first of these is larger than the second, the conclusion is that the expected value of the correlation coefficient is larger when computed from the means. It is, however, less significant. This is shown by the fact that the variance (V) will have the same expected value whether computed from B (or C) or from B + C, while the number of degrees of freedom on which V is based will obviously be greater if it is computed from B + C than if it is computed from B alone.

These conclusions are all valid for the expected values of the various quantities, but I have not been able to extend them to the results for a single sample.

This line of reasoning suggests, however, that if one had a regression equation computed from the means, it should be possible to improve the estimate of the regression coefficient -- or of its variance -- by a knowledge

of the intraclass variance of the dependent variable. The question I should like to ask Dr. Fisher is whether and how this can be done.

DR. FISHER: If we had the value C, one could, I suppose, reconstruct the regression coefficient based on the individual observations from the one based on the means and so modify your estimate, which, as you said, is unaffected by grouping the dependent variate. A better estimate of its error would also be obtained, and in applying a test of significance to the regression coefficient there would be a larger number of degrees of freedom. Supposing you took individual observations, one would calculate

$$\frac{S(Y(X - \bar{X}))}{S(X - \bar{X})^2},$$

as the estimate of the regression coefficient. There is absolutely no change if we multiply the parenthesis in the numerator by n_p and substitute $S(n_p \cdot \bar{Y} \cdot (X - \bar{X}))$, where \bar{Y} now stands for the mean of the n_p Y-values associated with the same value of X. S stands for summation of all individual observations, so that there can be no definite difference at all between the methods until you group the independent variate, and grouping the independent variate makes a certain change.

I think that is not related to this analysis at all, is it? What happens when you group the independent variate is perhaps most simply seen by saying that we are introducing errors of grouping into the value of X. The numerator will not be biased by errors of grouping. But some of the squares of deviations from the mean of the independent variate, if we introduce errors of any kind in the independent variate, will be inflated by them, and consequently one will get too low an estimate of regression whenever errors are introduced into the variate taken as the independent variable. That is a very troublesome thing to the physicist who may be dealing with interchangeable variates and cannot get away from the slight ambiguity in the choice between the line of X on Y, or the line of Y on X, but that is a different problem.

MR. FRIEDMAN: Dr. Fisher's conclusion is that the regression coefficient will be biased downwards by reason of the errors of grouping..

DR. FISHER: If you group the independent variate.

MR. FRIEDMAN: And that was the case I was thinking of.

DR. FISHER: You may say that the regression needs something like a correction for continuity, or a Sheppard's correction, which is essentially the same thing.

MR. FRIEDMAN: Has that correction been worked out so that it would be valuable?

DR. FISHER: Well, you could get an unbiased estimate of that denominator.

Supposing you group into interval H , the correction would be $H^2/12$, and you could deduct n times this expression from the apparent sum of squares. You would at any rate get an unbiased denominator. That is academic and theoretical, because, in fact, it is unlikely that errors of grouping are the only errors that the independent variable suffers from. An anthropologist measures people, let us say, to the closest inch, so that the error of grouping is a rectangular distribution of half an inch each way. But in fact, of course, one knows that the error of observation is quite likely to be as large, and so at the best, in observation and grouping like that, I attach little importance to Sheppard's correction, because it is a little academic.

MR. FRIEDMAN: But having obtained the unbiased estimate of B can the intraclass variance be used to improve or to get a better test of significance of the regression coefficient obtained from the means?

DR. FISHER: It seems to me that the variation within the array, which is C , only serves the purpose of increasing the number of degrees of freedom. Supposing you had only the means, you would then have to accept the test of significance based on the number of degrees of freedom available for B .

MR. STURGES: After our discussion yesterday* I was thinking that it might be of interest to approach the problem of the determination of the degrees of freedom in a time series correlation by mean of an analogous artificial series. A pair of such series may be gotten with dice by the following procedure:

- (a) Throw four dice; let the sum of spots be X_1 .

$$\text{Suppose we get } X_1 = \boxed{2} + \boxed{5} + \boxed{3} + \boxed{4} = 14$$

- (b) Throw two of them again, chosen at random; and leave the other two unchanged; call the new sum Y_1 .

$$\text{Suppose we get } Y_1 = \boxed{2} + \boxed{5} + \boxed{1} + \boxed{6} = 14$$

- (c) Return the dice to their positions for X_1 . Throw one die, chosen at random, leaving the other three unchanged; let the sum of spots be X_2 , for which let us suppose

$$X_2 = \boxed{2} + \boxed{4} + \boxed{3} + \boxed{4} = 13$$

- (d) Get Y_2 from these in the same way that Y_1 was gotten from X_1 .

- (c) Repeat, say to 100 pairs of X_i, Y_i .

Because the observed value of the S.D. of the X_i will be less than the true value, owing to the correlation of 0.75 between successive X_i , and because the same thing is true of the Y_i , the observed correlation

* See the report on yesterday's conference at the Cosmos Club.

between X and Y will exceed the true value 0.5. Since there are not 100 independent observations, the question arises, how many are there? The question might better be stated with reference to the S.D. of X only, not with regard to the correlation of X and Y.

DR. FISHER: If your mean is not independent, I don't see how it would affect the sigmas. Supposing you go on 40 times, and that none of the original dice is standing and your 40th time is independent of the first. Suppose you were to pick out every 40th trial, we still would not be free to take for our degrees of freedom the whole number of observables. If you took every 40th variant, you would get a distribution exactly the same as the original distribution, with a constant mean, say, of 14, and there are 40 such distributions, identical, no matter how much they may be related to each other, superimposed to make your whole distribution. They must have the same mean, or standard deviation, as if they went higher. What is introduced is a correlation between successive values.

MR. STURGES: That is exactly the point that I want to raise: That it is necessary to weed out these partial times in between and take just those. We are not free to pick everything we have in such a series but we must take some pains to insure independence.

DR. FISHER: You would need a longer series before you would get a good estimate of sigma.

MR. STURGES: Suppose I had two series of, say, 100 items each. In the first series each X is independent of every other X, but in the second series the situation will be different, for we shall suppose that it is derived by some system of chance like the one just described. Now the S. D. can be calculated from the first series, and there will be for it 99 degrees of freedom. But for the second series we are not permitted to say ... (interrupted by Dr. Fisher)

DR. FISHER: I have often wondered whether some set-up that would recognize the difficulties of time series would not be worth working out theoretically. Clearly we cannot equate the performance of the time series to any simple independent set of observations, but you might suppose that we took the first application that throws a biased or smooth curve, as the point of equilibrium of some observable point might be changed in the regular, determinable manner. What I am suggesting is not whether we can work out a summation of consequences this morning, but that it might be worth considering whether those consequences, if worked out, would be likely to reduce the characteristics of a time series in such a way as to make them more reliable and to make their actual significance expressible in terms of such a conclusion. How do you think that would go?

MR. STURGES: I think that some such device ought to be worked out, because most of the economists feel it is built up of three discrepant factors - the moving point of equilibrium, the random elements by the successive throws of dice, and there is the nexus between each observation and the next.

I think the tendency to take the fifteen series and say that we have 14 degrees of freedom has made a great many false inferences.

DR. FISHER: We have fourteen comparisons, but those comparisons are similar to each other. Whenever there is a general trend underlying, then the way I should put it is that the first few degrees of freedom would be heterogeneous from the others, just as they would be if one set of objects were in inches and another in a different measure; and of course, any three degrees of freedom is not necessarily heterogeneous with the other 11 degrees of freedom. There is no partial correlation of a with c when b is eliminated, so it simply is a combinational matrix to a considerable extent.

DR. SASULY: I hesitate to call attention to this from the point of an extended approach to the time series problem. I have done some work in that connection; but not much attention seems to have been paid to it. It is a problem that can be dealt with in this way: The people who make the academic observations thought they could make 14 observations in 14 years, or even 14 times 365, and that the number of degrees of freedom is common with the number of observations. That would be exceedingly indeterminate.

Some six or seven years ago it was necessary to assume that a time series involves, as in a hypothesis, the sort of thing we would meet in a physical experiment if one measured the distribution of current in a transformer of an electrical system. You can measure mechanical effects by an instrument and you can measure the current effects and have both recorded, and you can measure the temperature or the dynamic reaction, the attraction between the coils, and you can plot three curves. You have the same records that you would have if you considered the supply of some commodity and the price prevailing, and, let us say, the income. Getting the three records of these three variables and what your record would give you if you knew the price of eggs and the amounts taken in different markets, would be similar. You can see a certain connection between the three variables and you can get the usual Ohm's law for circuits in the electrical case and by the empirical criterion that it works, you can predict the distribution in similar circuits, and you will find the law verified. We don't quite expect to get the law of distribution of price, supply, and income. That is a much more difficult problem, but you do have something that is not quite what you would get out of the random series of dice.

I have been wanting to ask Dr. Fisher about this. Dr. Fisher did an extensive job around 1924* on the correlation of rainfall and yield. That is an ideal case of a mixture of variables that seems to be dynamical. Rainfall has to do with the motion of the earth around the sun on its axis, and you have random things, things that make successful and non-successful crops. Dr. Fisher carried that work a little further, namely, considered from the matter of seasonality the matter of how much valuable information one gets by eliminating a trend, as we say here, and dealing with the residuals. Personally, I like to deal with this problem by considering the trend itself, and I don't feel competent to handle these random things.

* Phi. Trans. Royal Soc. B213, 89-142 (1924)

Now, to what extent is this valid? The prevailing view would be that if one did a job like that, it would not be reliable; for you have not the coefficient of reliability and the coefficient of dependability that one must produce.

What I want to bring out is this: Those who are most skilled in experience in handling random effects could show us how much error to introduce or that one could ignore in the consideration of random effects. Concretely, I would take this case: Suppose you had data like that (pointing to a time series on the blackboard) and you don't fit them by least squares at all. Let us say that this involves, speaking from experience, the matter of construction, the amount of housing, and the amount of building units over a time; and the most plausible curve you were able to find is one that was fitted by the following sort of criterion. I would get the curve to be such that the area under the curve would be equivalent in each strip to the area that you got by summing up the data considered as rectangles. It is not a least squares criterion. We have those constants, but there is no way of expressing the usual measures of probable error and variance.

DR. FISHER: Do you feel there is any advantage in that method of fitting? It is difficult to discuss unless you feel there is some advantage.

DR. SASULY: The problem as we set it up in this concrete case shows the difficulty of testing how building structures during the depression was determined by certain factors, like labor costs, which was a hot question; the other was the question of credit. By the first set-up, the standard least squares set-up, the solution we got was a little worse than the one that Frisch made. There was a question of how building was connected with credit, and the result was opposite what it should have been. We know that the more stringent conditions are, the less building we have.

DR. FISHER: We know that if the builder pays higher rates of interest, he is deterred from building. But we don't know that higher rates of interest are inevitable to the demands for building.

DR. SASULY: It was something a little more definite. It was in effect a measure of credit, something to do with a measure of foreclosures. The point is that when we made the other postulate of fitting this curve by equivalent strips, then the coefficient came out plausible. By postulating the equivalent of least squares, we found one region during the war where there were a lot of conditions that did not fit into this picture; the discrepancies were pretty large, and the criterion of least squares gave this region greater weight for determining the coefficient. That is an example where we find a much wider approach is necessary in practical problems than the standard of least squares, so we have this extra justification.

DR. FISHER: I do feel there is a very great advantage in sticking to either the least squares or, if there were any alternative, to some widely recognized method of fitting merely from the fact that one can go and find out from another man what he has been doing; and second, that the consequences of the least squares, as to the residual squares, have been

much more fully worked out. Scarcely a beginning has been made in the working out of any other method, and so I think that it is desirable, where possible, to conform to the original line.

On the other hand, the first point you made was an extremely good one -- namely, that when we fit curves to time series, we ought to confine them to what we are doing, and merely eliminate the trend; and I should like to suggest that we ought to think of any such process as supporting the different components of the data. There are different types of facts, and from that point of view, I think one can recognize some of the main features.

One has, as an example in polynominal fitting, my own work, which you were kind enough to mention -- a case where the error related to wheat yields from 1853 to 1918, on the same series of plots. Mere common sense and historical knowledge suggest that very probably the wheat crop in the 50's was not in any way comparable with the wheat crop in the present century. Tractors are now being used; but I found, interesting enough, that oxen were still used at the beginning of the series. About 1880 or 1884, a series of educational acts were passed which had the effect of withdrawing population from agricultural occupations, with the reasonable consequence in that particular field that wheat infestations had grown up rather rapidly; and so there were a few things that we could see, but we should overestimate our intuitional knowledge very greatly if we could say there were 300 or so errors which we could see.

There was a case of elimination. If the yields had gone on like that (indicating the linear upward trend of the first part of the series) -- well, a straight line would have taken out all of those concerned and which we wished to eliminate. Actually, that would be more the figure that I was faced with, and although some of the slow change was reasonable and ascribable, holding to real differences in the rainfall, yet it was not sufficiently safe to use it to allow these heavy differences to weigh in and bias my opinion as to what the rainfall was really doing. What I wanted to rely on in judging the effect of the rainfall was to compare that with the yield in successive years, in which the general treatment of the land and the condition of the soil and the technique of farming would be much more comparable than it would be in comparisons in point of time.

Editors' note: For further discussion of this subject the reader is referred to the following papers: M.S. Bartlett, Journal of the Royal Stat. Soc. 98, 536-543 (1935); C.F. Roos, Econometrica 4, 368-381 (Oct. 1936); G. Udney Yule, Journal of the Royal Stat. Soc. 84, 497-537 (1921).

MR. STAUBER: I should like to ask about the distribution of an inquiry over a geographical area for the purpose of determining information concerning some one characteristic. Let us suppose that there is an area, say a state, perhaps the United States. On the basis of previous work, we have been able to divide that area into a number of regions which are more homogeneous than the larger area but yet not completely homogeneous.

Assume that there are to be a number of schedules or inquiries, or a number of farms to be interviewed, and that such number had already been determined previously. We may proceed with the sampling in a number of ways; for example, assume that all of the farms have been listed, you may select every 20th one from an alphabetical series, or in some such fashion you could proceed to select one out of every 20 farms. That would be one procedure and would probably be equivalent to a random selection.

Another procedure that might be adopted is to assume, on the basis of our previous knowledge, that each of the regions you might select may be subdivided into smaller sub-regions, which the investigator believes to be representative of the region, but schedules are sent to all the farms in the selected subregions. I should like some comments as to the relative merits of those two procedures, assuming the same number of schedules to be sent, or farms to be interviewed, in each situation.

DR. FISHER: Personally, I much prefer the first program, merely for the simple reason that it does fulfill this condition of sampling, which seems a good one so that every unit has an equally good chance of appearing in your sample, whereas that sub-division into typical regions, with the rejection of other regions, means that no farm in a rejected region has a chance of getting in the sample. As the end point of view is strongly in the mind of the man who makes the selection, the sample cannot be genuinely representative.

With respect to choosing at random, I don't think there is any lack of adequate methods of choosing at random from the list. One need not tie oneself down to taking every 20th house down a street or every 20th farm in the land. One should, I think, always carefully subdivide his selections, the first geographically. I suppose you are thinking of such broad regions as the cotton region, or the corn region, but even within those regions certain divisions are reasonable and possible, such as the size of farm, proportion arable, etc. There may be other divisions which may or may not be worth noticing.

Suppose we make one sub-strata according to the size of farm and another according to the proportion arable, and then select one in 20 from each substrata and at random. We have, then, reconciled the genuine representativeness with the consideration of stratification, in that our sample error will not only show a difference which we may attribute to the excess of any farms of the same size and the same proportion, but which would be clearly less than the difference between farms in general.

So I think there is a great deal to be said for giving a great deal of time to our stratifications of the whole on the basis of what is known about it prior to an entirely objective, random sample, by which I mean something carried out by means of some one or other of the ideas which the ingenuity of gamblers has developed, like the throwing of dice or the shuffling of cards or books of numbers.

MR. FRIEDMAN: If the number of cases that can be obtained is limited, and if the weights, i.e., the proportion of farms in each geographical area, are known with accuracy, say, given by the Census, would it not be desirable to distribute the cases equally among the sub-classes?

DR. FISHER: You know the number of farms of 10 to 50, 50 to 100, 100 to 200 acres, etc., and then you are asking whether you should take the same proportion from each class.

MR. FRIEDMAN: Would it not be preferable to take the same number from each class rather than the same proportion?

DR. FISHER: I think the same proportion. What advantage have you in mind in taking the same number? There would be a few adjoining ranches which might be in a class by themselves.

MR. FRIEDMAN: Assuming you have a fixed total number to distribute, by taking the same number in each class, it would seem to me that the final standard error of your weighted average would be somewhat reduced, and in particular, that the standard error of your regression coefficients would be reduced.

DR. FISHER: Suppose you represented the number of cases included in the i -th sub-class of a series of stratified samples by n_i , the variance within that sub-class by V_i , and the actual number in each sub-class by N_i . The variance of the weighted mean multiplied by the actual total number in your register, is

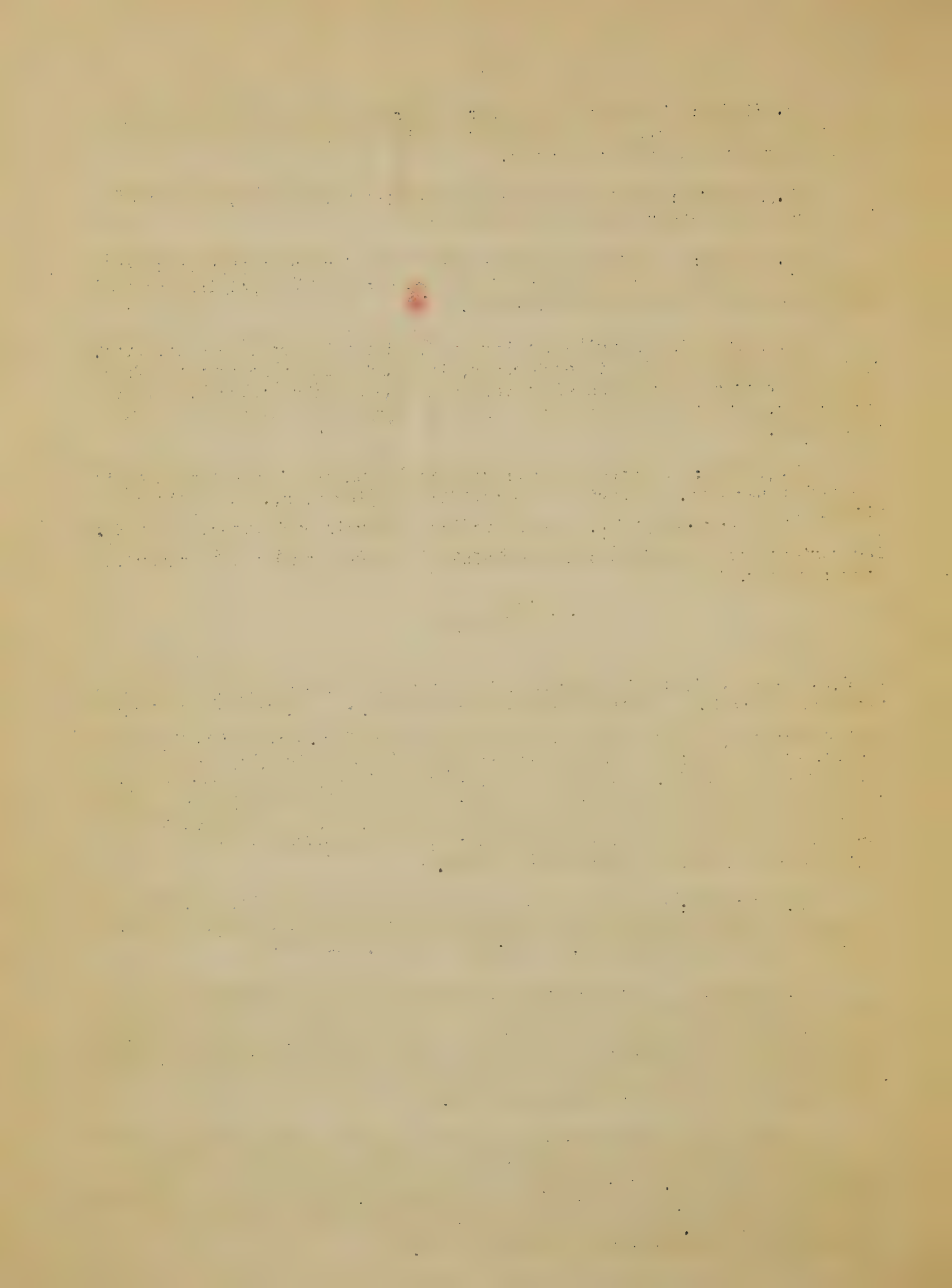
$$S (N_i^2 V_i / n_i)$$

If V is the same for each sub-class, I think that you would have minimum variance if the n_i were in proportion to the N_i . But if you found from your sample some sub-classes for which the variance was greater or less than others, presumably you would have to sample more extensively where you have the greater variances. It would be worthwhile reducing the larger contributions to your total variance in high proportion at the expense of letting the smaller contributions become somewhat unfavorable. If these variances were constant for the different strata, I think the error would be minimized by taking a proportionate sample.

MR. FRIEDMAN: If you were to take an equal number in each class instead of a proportionate number, wouldn't your regression equation be more accurate in that the scatter of the X_i would be increased?

DR. FISHER: If you have an independent variate varying from 10 to 90, and if the matter of primary interest is the regression of something on that percentage, then with a limited number of cases, you don't want to waste them by having them in the middle, and you may be arbitrary by having a lot at the two ends, if you are sure how the thing is laid out. You may lose some information by omitting the ends.

There are so many purposes to which a sample scheme may be put that it is almost impossible to conceive of a sampling system that should be ideal for all of them. Personally, I should almost be tempted to sample equally, at any rate, at the first stage, perhaps with a view to correcting your deficiencies by more intensive sampling at a second stage if any questions of importance still remain in doubt.



DEC 9 - 1938

